A THEORY OF SOME MULTIPLE DECISION PROBLEMS*. II

By E. L. Lehmann[1]

University of California, Berkeley

Summary. The theory of Part I is extended to problems in which it is permitted not to come to a definite conclusion regarding one or more of the questions under consideration. Some problems are also investigated in which, from a single set of observations, one wishes to answer a number of questions in sequence. Here the nature of the question at a later stage will depend on the answers obtained at the earlier stages.

9. Decision procedures permitting partial conclusions. It is a common feature of all the problems treated in Part I, that a fixed partition of the parameter space Ω into sets Ω_i is given, and that one wishes to determine which of these sets contains the true parameter point. There are however many statistical problems, such as the estimation by confidence sets, in which the possible decisions do not correspond to the sets of a fixed partition. In particular, this is the case in the field of statistical inference, when the statistician is free to decide how sharp a statement he can reliably make on the basis of the observations. We shall show in the present section how such problems may be generated by the simultaneous consideration of a number of two-decision problems as in Part I, if one suitably modifies the interpretation of the decisions involved.

Previously we were concerned with testing a set of hypotheses $H_\gamma: \theta \varepsilon \omega_\gamma$, so that in each component problem the choice lay between the two decisions $\theta \varepsilon \omega_\gamma$ (acceptance of H_γ) and $\theta \varepsilon \omega_\gamma^{-1}$ (rejection of H_γ). Suppose now instead that the statistician is asked only whether the data reject the hypothesis, and that in case they do not, no alternative positive statement is required. The choice may then be said to lie between the statements $\theta \varepsilon \omega_\gamma^{-1}$ and $\theta \varepsilon \omega_\gamma^0 = \Omega$.

This actually appears to be the point of view taken by Duncan ("Multiple range and multiple F tests", *Biometrics*, Vol. 11 (1955), pp. 1–42) in his formulation of this class of multiple decision problems to which reference is made in section 1 of the present paper.

If one considers simultaneously a number of such problems, one is faced with a multiple decision problem in which the different possible decisions correspond to the statements that a certain number of the hypotheses H_γ are false, but where nothing is said regarding the remaining hypotheses. This is equivalent to the statement that the parameter point θ lies in the set

$$\Omega_i = \bigcap_\gamma \omega_\gamma^{y_i} \tag{9.1}$$

Received October 12, 1956.

* To rectify a printing error in Part I, in the March, 1957 issue of these *Annals* replace page 14, formula (4.12) through page 17, line 23 by page 70, line 8 through page 72, next-to-last line. Ed.

[1] Work done while the author was a Fellow of the John Simon Guggenheim Memorial Foundation.

where the y's are -1 for each rejected H_γ and 0 for the others. In the particular case that all of the y's are zero, we have $\Omega_i = \Omega$, and thus make no statement whatever about the position of θ. As before, it may of course happen that some of the formal intersections (9.1) are empty, and we shall then restrict the Ω's to denote the nonempty ones and require that none of the possible decisions should correspond to these empty intersections. If we assume that in the simultaneous consideration of a number of problems the losses are additive so that the total loss is the sum (or a weighted sum) of the losses of the component problems, and if we suppose the losses to be a_γ for rejecting H_γ when $\theta \varepsilon \omega_\gamma$, b_γ for not rejecting it when $\theta \varepsilon \omega_\gamma^{-1}$, and zero in the other two cases, the total loss is again given by (2.2) and (2.3) with

$$x_{i\gamma} = 2y_{i\gamma} + 1. \tag{9.2}$$

Since in particular the loss function of the basic two-decision problem is unchanged, the associated optimum unbiased procedure of Part I will retain its optimum property in spite of the reinterpretation of one of the decisions. It now leads to the statements $\theta \varepsilon \omega_\gamma^{-1}$ and $\theta \varepsilon \omega_\gamma^0$ as $X \varepsilon A_\gamma^{-1}$ or $X \varepsilon A_\gamma$ where A_γ is the acceptance region of the best unbiased test of H_γ. The simultaneous carrying out of a number of these tests then still leads to a procedure in which the decision $d_i : \theta \varepsilon \Omega_i$ is taken when X falls in the set

$$D_i = \bigcap_\gamma A^{x_i} \tag{9.3}$$

but where Ω_i is now defined by (9.1) and (9.2) instead of (2.1).

As has already been pointed out the sets Ω_i, which define the possible decisions, no longer constitute a partition of the parameter space. Instead, they are generated through intersections from the class $\{\omega_\gamma^{-1}, \gamma \varepsilon \Gamma\}$, that is, they constitute the smallest class that is closed under intersections and contains the sets ω_γ^{-1}. It may happen that two of these Ω's are equal, $\Omega_i = \Omega_j$, say, and one would then wish to identify the associated decisions. On the other hand, viewing the problem as a product one must consider all of the formal intersections (9.1) as distinct. Otherwise the definition of the loss function, for example, would become ambiguous since the losses resulting from decisions d_i and d_j when θ is in some Ω_k would usually not be the same even though $\Omega_i = \Omega_j$. Fortunately, the difficulty arises in the applications we wish to make only in cases in which it can be overcome by a natural further restriction of the decision space. Suppose namely that H_γ and H_δ are two hypotheses with

$$\omega_\gamma^{-1} \subset \omega_\delta^{-1}$$

so that the two intersections $\omega_\gamma^{-1} \cap \omega_\delta^{-1}$ and $\omega_\gamma^{-1} \cap \omega_\delta^0$ are identical. It then seems reasonable that whenever the data lead to the rejection of H_γ one would also wish to reject the more restrictive hypothesis H_δ. (In part I this was actually part of the compatibility requirement.) With such tests the decisions $\theta \varepsilon \omega_\gamma^{-1} \cap \omega_\delta^0$ would never be reached, and the conflict would thus be avoided. For this reason we shall eliminate from the list of permissible decisions not only the formal intersections

(9.1) that are empty, but also those for which the intersection Ω_i equals some other Ω_j satisfying $y_{j\gamma} \leqq y_{i\gamma}$ for all γ. It follows from the general discussion of Sec. 7 that (9.3) defines a 1:1 correspondence between families of tests which are compatible with this set of restrictions, and decision procedures for the restricted product problem. It is useful to note further that for these restrictions the formation of restricted products of decision problems also retains the property of being commutative and associative, which it obviously had in the problems of Part I.

Let $\mathcal{P}_+(\theta_0)$ denote the two-decision problem discussed above, in which the choice lies between the statements $\theta > \theta_0$ and $-\infty < \theta < \infty$, and let $\mathcal{P}_-(\theta_0)$ denote the dual problem in which the first of these possibilities is replaced by the statement $\theta < \theta_0$. By considering these two problems simultaneously one obtains a three-decision problem with loss function

(9.4)

	d_2 $\theta < \theta_0$	d_0 $-\infty < \theta < \infty$	d_1 $\theta > \theta_0$
$\theta < \theta_0$	0	b	$a + b$
$\theta = \theta_0$	a	0	a
$\theta > \theta_0$	$a + b$	b	0

As an example suppose that θ measures the difference in quality of two products which are being compared by an impartial research organization. The decisions d_1 and d_2 claim superiority for one or the other of the products, while d_0 states that the data are inconclusive and that neither of the two products can be ascertained to be better than the other. It is an advantage of such a formulation over the more conventional one in which d_0 is replaced by the statement $\theta = \theta_0$, that it enables the statistician to control the probability of error. In the standard situation with D_2 and D_1 given by

$$D_2: T \leqq C_1 \quad \text{and} \quad D_1: T \geqq C_1' \tag{9.5}$$

where $P_\theta\{T \leqq C_1\}$ and $P_\theta\{T \geqq C_2'\}$ are $< \alpha$ for $\theta > \theta_0$ and $\theta < \theta_0$ respectively, and where $P_{\theta_0}\{T \leqq C_1\} = P_{\theta_0}\{T \geqq C_2'\} = \alpha$, the maximum probability of error occurs when $\theta = \theta_0$, and is 2α. (A very similar formulation was discussed by Bahadur, "A property of the t-statistic," *Sankhya*, Vol. 12 (1952), pp. 79–88.)

The loss function (9.4) is not appropriate in situations in which a definite decision is preferred to d_0 even when $\theta = \theta_0$. A formulation which is more suitable for this case is obtained if in the one-sided problems $\mathcal{P}_-(\theta_0)$ and $\mathcal{P}_+(\theta_0)$ one replaces the decisions $\theta < \theta_0$ and $\theta > \theta_0$ by $\theta \leqq \theta_0$ and $\theta \geqq \theta_0$ respectively, so that these two component problems are given by

(9.6a)

	$-\infty < \theta < \infty$	$\theta \leqq \theta_0$
$\theta > \theta_0$	0	a
$\theta \leqq \theta_0$	b	0

and $(b < a)$

(9.6b)

	$-\infty < \theta < \infty$	$\theta \geqq \theta_0$
$\theta < \theta_0$	0	a
$\theta \geqq \theta_0$	b	0

The simultaneous consideration of these two problems leads to a four-decision problem with loss table

(9.7)

	d_2 $\theta \leqq \theta_0$	d_0 $-\infty < \theta < \infty$	d_1 $\theta \geqq \theta_0$	d_3 $\theta = \theta_0$
$\theta < \theta_0$	0	b	$a + b$	a
$\theta = \theta_0$	b	$2b$	b	0
$\theta > \theta_0$	$a + b$	b	0	a

It turns out that this formulation leads to essentially the same solution as the previous one, with D_2 and D_1 given by (9.5), decision d_0 being taken when $C_1 < T < C_1'$, and decision d_3 not occurring at all.

We mention finally that still another problem leads to the same solution, namely that given by

(9.8)

	d_2 $\theta \leqq \theta_0$	d_0 $-\infty < \theta < \infty$	d_1 $\theta \geqq \theta_0$	
$\theta < \theta_0$	0	b	a'	$\begin{pmatrix} b < a' \\ b < b' \end{pmatrix}$.
$\theta = \theta_0$	0	b'	0	
$\theta > \theta_0$	a'	b	0	

The level α of (9.5) is in this case given by

$$\alpha = \frac{b' - b}{a' + 2(b' - b)}.$$

10. Partial classification of one or more parameters. (i) Let θ be a real parameter and suppose that we wish to determine, as far as possible, its position relative to two given values $\theta_1 < \theta_2$. A procedure may be generated by considering simultaneously the four problems $\mathcal{P}_{\pm}(\theta_i)$, $i = 1, 2$. The resulting problem offers the choice between the decisions

(10.1)

$$\begin{aligned}
&d_1: \{\theta < \theta_1\} = \{\theta < \theta_1\} \cap \{\theta < \theta_2\} \\
&d_2: \{\theta < \theta_2\} = \{-\infty < \theta < \infty\} \cap \{\theta < \theta_2\} \\
&d_3: \{\theta_1 < \theta < \theta_2\} = \{\theta_1 < \theta\} \cap \{\theta < \theta_2\} \\
&d_4: \{\theta > \theta_1\} = \{\theta > \theta_1\} \cap \{-\infty < \theta < \infty\} \\
&d_5: \{\theta > \theta_2\} = \{\theta > \theta_2\} \cap \{\theta > \theta_1\} \\
&d_6: \{-\infty < \theta < \infty\} = \{-\infty < \theta < \infty\} \cap \{-\infty < -\theta < \infty\}.
\end{aligned}$$

Here the sets $\{\theta < \theta_1\}$ and $\{\theta > \theta_2\}$ may also be represented as the intersections $\{\theta < \theta_1\} \cap \{-\infty < \theta < \infty\}$ and $\{\theta > \theta_2\} \cap \{-\infty < \theta < \infty\}$. However these decisions are ruled out by the convention of the previous section.

Suppose now that the tests of the four generating hypotheses have rejection regions

$$T \leqq C_1, \qquad T \leqq C_1', \qquad T \geqq C_2, \qquad T \geqq C_2'$$

for $H{:}\theta \leqq \theta_1$, $\theta \geqq \theta_1$, $\theta \leqq \theta_2$, $\theta \geqq \theta_2$ respectively, where the constants are determined by

$$P_{\theta_1}\{T \leqq C_1\} = P_{\theta_1}\{T \geqq C_1'\} = P_{\theta_2}\{T \leqq C_2\} = P_{\theta_2}\{T \geqq C_2'\} = \alpha.$$

Compatibility requires that the intersections

$$\{T \leqq C_i\} \cap \{T \geqq C_i'\}, \qquad \{T \leqq C_1\} \cap \{C_2 < T < C_2'\}$$
$$\text{and } \{C_1 < T < C_i'\} \cap \{T \geqq C_2'\}$$

should be empty, and hence that

$$C_1 < C_1', \qquad C_2 < C_2', \qquad C_1 < C_2 \text{ and } C_1' < C_2'.$$

These conditions are satisfied if $\alpha < \frac{1}{2}$, and if for each fixed C

$$P_\theta\{T \leqq C\} > P_{\theta'}\{T \leqq C\} \quad \text{when} \quad \theta < \theta'. \tag{10.2}$$

According to (9.3) the resulting procedure is given by

$$\begin{aligned} &D_1{:}T \leqq C_1 && D_4{:}\max(C_1', C_2) \lesseqgtr T < C_2' \\ &D_2{:}C_1 \lesseqgtr T < \min(C_1', C_2) && D_5{:}T \geqq C_2' \\ &D_3{:}C_1' \leqq T \leqq C_2 && D_6{:}C_2 < T < C_1' \end{aligned} \tag{10.3}$$

where $\lesseqgtr$ is $<$ when $C_1' < C_2$ and $\leqq$ when $C_1' \geqq C_2$. Depending on the sign of the difference $C_2 - C_1'$ and hence on the distance between θ_1 and θ_2, only one of the decisions d_2 and d_6 will occur. For intermediate values of T the positive statement $\theta_1 < \theta < \theta_2$ will be made only if θ_1 and θ_2 are not too close. Otherwise such T-values will leave the position of θ in doubt.

If (10.3) holds, the probability of the procedure leading to a false statement never exceeds 2α. For the probability of error is equal to

$$P_\theta\{T \geqq C_1'\} \leqq P_{\theta_1}\{T \geqq C_1'\} = \alpha \qquad \text{for} \quad \theta < \theta_1$$

$$P_{\theta_1}\{T \leqq C_1\} + P_{\theta_1}\{T \geqq C_1'\} = 2\alpha \qquad \text{for} \quad \theta = \theta_1$$

$$P_\theta\{T \leqq C_1\} + P_\theta\{T \geqq C_2\} \leqq P_{\theta_1}\{T \leqq C_1\} + P_{\theta_2}\{T \geqq C_2\} = 2\alpha$$
$$\text{for} \quad \theta_1 < \theta < \theta_2,$$

and similarly for $\theta \geqq \theta_2$.

In the usual applications T is a function of a sample, and as the sample size increases T tends to θ in probability. The procedure is then consistent in the

sense that

$$P(D_i) \to 1 \quad \text{for} \quad \theta \in \Omega_i \qquad (i = 1, 2, 3).$$

Except when θ is exactly equal to θ_1 or θ_2, the probability is therefore 1 that one of the three sharp statements d_1, d_2, or d_3 will be made, and that this statement will be correct. Since

$$P_{\theta_1}(D_2) = P_{\theta_2}(D_4) = 1 - 2\alpha,$$

it also follows that

$$P_{\theta_1}(D_3) = P_{\theta_2}(D_4) \to 1$$

if one lets α tend to zero as n tends to infinity.

A slightly different procedure results for problem (10.1) when one is concerned with a sample $X_1, \cdots, X_n$ from $N(\theta, \sigma^2)$. Here the tests of the four generating hypotheses have rejection regions

$$(\bar{X} - \theta_1)/S \leqq -C, \qquad (\bar{X} - \theta_1)/S \geqq C,$$
$$(\bar{X} - \theta_2)/S \leqq -C, \qquad (\bar{X} - \theta_2)/S \geqq C,$$

and the induced multiple decision procedure is given by (10.3) with

$$T = \bar{X}/S, \qquad C_i = -C + \frac{\theta_i}{S}, \qquad C_i' = C + \frac{\theta_i}{S}. \tag{10.4}$$

Both of the decisions d_2 and d_6 occur in this procedure with positive probability, d_2 only in cases in which $S \leqq (\theta_2 - \theta_1)/C$ and d_6 only when the opposite inequality holds. The remarks concerning error control and consistency require no change.

As an application consider the comparison of two normal populations $N(\xi, \sigma^2)$ and $N(\eta, \sigma^2)$ on the basis of samples $X_1, \cdots, X_m$ and $Y_1, \cdots, Y_n$. With $\theta = \eta - \xi$, $\theta_1 = -\Delta$, $\theta_2 = \Delta$, the possible decisions become

$$d_1: \eta - \xi < -\Delta, \qquad d_3: -\Delta < \eta - \xi < \Delta, \qquad d_5: \eta - \xi > \Delta,$$
$$d_2: \eta - \xi < \Delta, \qquad d_4: \eta - \xi > -\Delta, \qquad d_6: -\infty < \eta - \xi < \infty.$$

Here d_1 states that ξ is significantly larger than η, d_2 that η is not significantly larger than ξ, d_3 that the two means do not differ significantly, etc. The procedure is given by (10.3) with $T = (\bar{Y} - \bar{X})/S$ and

$$C_1 = -C - \frac{\Delta}{S}, \qquad C_2 = -C + \frac{\Delta}{S}, \qquad C_1' = C - \frac{\Delta}{S}, \qquad C_2' = C + \frac{\Delta}{S}.$$

Problem (10.1) leads to still another type of procedure in the case of two independent Poisson variables, say X and Y, where one wishes to classify the ratio $\rho = \lambda/\mu$ of the parameters of the two distributions with respect to two values $\rho_1 < \rho_2$. Here the tests of the four generating hypotheses $\rho \leqq \rho_i$, $\rho \geqq \rho_i (i = 1, 2)$ are carried out conditionally, given the value of $X + Y$. The

conditional distribution of Y given $X + Y = m$ is the binomial distribution $b(p, m)$ corresponding to m trials and probability $p = \mu/(\lambda + \mu)$ of success. The conditional situation is therefore of the kind discussed at the beginning of this section, and leads to the procedure (10.3) with $T = Y$ and $\theta = p$.

Whether d_3 or d_6 occurs depends here on m, d_3 being associated with large values of m. To see this, let $F_{p,m}(t)$ denote the conditional cumulative distribution function, given $X + Y = m$, of $Y + U$ where U is independent of Y and is uniformly distributed on (0, 1). Then $C_1' = C'(m)$ and $C_2 = C_2(m)$ are determined by

$$F_{p_2,m}(C_2) = 1 - F_{p_1,m}(C_1') = \alpha. \tag{10.5}$$

We shall show that $C_1'(m_1) < C_2(m_1)$ implies that also $C_1'(m_2) < C_2(m_2)$ for all $m_2 > m_1$. Since $F_{p_2,m}(C_2) = \alpha$, $F_{p_1,m}(C_2)$ is the power of the most powerful level α test for testing p_2 against p_1 in a binomial distribution $b(p, m)$. If $C_1'(m_1) < C_2(m_1)$ it follows from (10.9) that in the case of m trials this power is greater than $1 - \alpha$; but then it must also exceed $1 - \alpha$ for $m_2 > m_1$ trials so that

$$F_{p_1,m_2}(C_1') = 1 - \alpha < F_{p_1,m_2}(C_2),$$

and hence $C_1'(m_2) < C_2(m_2)$.

It is interesting to note that in all of these problems the choice between decision d_3 and d_6 depends on the distance between θ_1 and θ_2 relative to the amount of information the data contain for the problem.

While the classification of θ with respect to a single value θ_0, or two values $\theta_1 < \theta_2$, are the most interesting cases, let us consider briefly also the problem of classifying θ with respect to a countable set of values $\cdots < \theta_{-2} < \theta_{-1} < \theta_0 < \theta_1 < \cdots$. This is generated by the problems $\mathcal{P}_{\pm}(\theta_i)$, $i = 0, \pm 1, \pm 2, \cdots$. Supposing that $\theta_i \to \pm\infty$ as $i \to \pm\infty$ and letting $\theta_{-\infty} = -\infty$, $\theta_\infty = \infty$, the possible decisions consist of the totality of statements $\theta_i < \theta < \theta_j$. The decision $\theta_i < \theta < \theta_j$ corresponds to the individual decisions that $\theta < \theta_k$ for $k \geqq j$ and $\theta > \theta_k$ for $k \leqq i$, and that the position of θ is left in doubt with respect to the points θ_k with $i < k < j$.

The limiting case of this problem, in which one wishes to classify θ with respect to all possible values of θ_0, is obtained by considering simultaneously the problems $\mathcal{P}_{\pm}(\theta_0)$ for all θ_0. The possible decisions then consist of the totality of statements $\underline{\theta} < \theta < \bar{\theta}$, and if $a_\gamma = a$, $b_\gamma = b$ for all γ, (9.3) yields precisely the standard confidence intervals for θ with confidence coefficient $1 - 2\alpha$. The loss function resulting from the additivity assumption is in the simplest case[2]

$$\begin{cases} (a + b)(\underline{\theta} - \theta) + b(\bar{\theta} - \underline{\theta}) & \text{if } \theta < \underline{\theta} \\ b(\bar{\theta} - \underline{\theta}) & \text{if } \underline{\theta} < \theta < \bar{\theta} \\ (a + b)(\theta - \bar{\theta}) + b(\bar{\theta} - \underline{\theta}) & \text{if } \theta > \bar{\theta}. \end{cases} \tag{10.6}$$

[2] A loss function of this type was suggested by Wolfowitz [4].

More generally one may replace $\bar{\theta} - \underline{\theta}$ by $\int_{\underline{\theta}}^{\bar{\theta}} d\mu(\theta)$, and similarly for the lengths of the other intervals. Unfortunately the condition of unbiasedness becomes very difficult to interpret in the present problem, and it is doubtful that for the standard distributions the procedure is unbiased with uniformly minimum risk, as is the case with the other problems of this section.

It should be pointed out that all these procedures concerning a single real-valued parameter θ can be obtained from the standard confidence intervals concerning θ. When classifying θ with respect to θ_1 and θ_2 for example, one can state $\theta < \theta_1$ if $\bar{\theta} < \theta_1$, and $\theta_1 < \theta < \theta_2$ if $\theta_1 < \underline{\theta} < \bar{\theta} < \theta_2$. On the other hand, if $\underline{\theta} < \theta_1 < \bar{\theta} < \theta_2$, only the conclusion $\theta < \theta_2$ is possible, the relation of θ to θ_1 being left in doubt. This approach, however, does not yield any optimum properties for these procedures, and does in fact not carry over to problems involving more than one parameter.

In Example (iii) of Sec. 3 we considered the comparison of a number of normal populations $N(\theta_i, \sigma^2)$. The consistency difficulties that occurred in combining the decisions $\theta_i \leqq \theta_j$ for different pairs (i, j) disappear if one treats the problem from the present point of view, which is exactly that from which the problem was treated by Duncan, "Multiple range and multiple F-tests," *Biometrics*, Vol. 11 (1955), pp. 1–42. For each pair (θ_i, θ_j) the possible decisions are now $\theta_i < \theta_j$, $\theta_i > \theta_j$, and $-\infty < \theta_j - \theta_i < \infty$ instead of the earlier $\theta_i = \theta_j$. Since there is no loss in omitting the vacuous statements, the total decisions consist in the ordering of some but not necessarily all of the pairs (θ_i, θ_j). In the case of three populations, for example, the following decision types will occur:

(a) $\theta_i < \theta_j < \theta_k$, (c) $\theta_i < \theta_j$, $\theta_i < \theta_k$

(b) $\theta_i < \theta_k$, $\theta_j < \theta_k$, (d) $\theta_i < \theta_j$

(e) no statement.

We shall now show that for this procedure the probability of error can be controlled through the choice of α, and that its maximum is in fact attained when all of the θ's are equal and is then given by

$$P\left\{\frac{|\bar{X}_j - \bar{X}_i|}{S} > C_{ij} \quad \text{for some} \quad i, j\right\}. \tag{10.7}$$

In the particular case of equal sample sizes this becomes

$$P\left\{\frac{\max |\bar{X}_j - \bar{X}_i|}{S} > C\right\} \tag{10.8}$$

where C is the cut-off point of the one-sided t-test at level α.[3] The probability of error is the probability measure of the set

$$\cup\left\{\frac{\bar{X}_j - \bar{X}_i}{S} > C_{ij}\right\} + \cup\left\{\frac{\bar{X}_\ell - \bar{X}_k}{S} < C_{k\ell}\right\} \tag{10.9}$$

[3] For a table of the values C for which this probability is 1 per cent or 5 per cent see [2], where also a number of related tables are discussed.

where the first union is extended over all pairs (i, j) for which $\theta_j \leqq \theta_i$ and the second union over all pairs (k, ℓ) for which $\theta_\ell \geqq \theta_k$. Let $X_i^* = X_i - \theta_i$, so that the X_i^* are distributed as $N(0, \sigma^2)$. Then $\theta_j \geqq \theta_i$ and $\bar{X}_j - \bar{X}_i > C_{ij}S$ imply $X_j^* - X_i^* > C_{ij}S$, and $\theta_\ell \leqq \theta_k$ and $\bar{X}_\ell - \bar{X}_k < C_{k\ell}S$ imply $X_\ell^* - X_k^* < C_{k\ell}S$. The probability of the set (10.9) is therefore not decreased if one replaces the X's by X^*'s, nor if one extends the union over all pairs (i, j) and (k, ℓ). But this is equivalent to evaluating the probability of (10.9) under the assumption that all of the θ's are equal, which completes the proof.

11. Decision problems with simple loss functions. As a tool for proving the procedures of Secs. 9 and 10 to be unbiased with uniformly minimum risk, we shall now give an extension of Theorem 2 (Sec. 7), which is valid for a rather general class of decision problems. We shall say that a decision problem $\mathcal{P}$ is *simple* if it satisfies the following two conditions.

(a) Its loss function $W(\theta, d)$, considered for fixed d as a function of θ, has sets of constancy independent of d, that is, there exists a partition Π of the parameter space Ω into sets Θ_i, $i \in I$, such that $W(\theta, d)$ is independent of θ on each Θ_i. We may then write

$$W(\theta, d) = V_i(d) \quad \text{for} \quad \theta \in \Theta_i . \tag{11.1}$$

(b) With respect to some convergence notion in Ω, $\theta_n \to \theta_0$ implies $E_{\theta_n}\psi(X) \to E_{\theta_0}\psi(X)$ for each integrable ψ or, if all of the functions V_i are bounded, for each bounded ψ.

We shall require the following properties of simple decision problems.

(i) *For any procedure δ the risk function $R_\delta(\theta)$ is continuous on each set of the partition* Π. The risk function is given by

$$R_\delta(\theta) = E_\theta V_i[\delta(X)] \quad \text{for} \quad \theta \in \Theta_i . \tag{11.2}$$

Hence θ_n, $\theta_0 \in \Theta_i$ and $\theta_n \to \theta_0$ imply

$$R_\delta(\theta_n) = E_{\theta_n} V_i[\delta(X)] \to E_{\theta_0} V_i[\delta(X)] = R_\delta(\theta_0),$$

as was to be proved.

(ii) *Unbiasedness of a procedure δ implies the continuity of its risk function.* By (i) it is enough to prove this for boundary points of Π. Let θ_0 be such a boundary point, and suppose that $\theta_0 \in \Theta_i$ and $\theta_0 = \lim_{n\to\infty} \theta_n$ with $\theta_n \in \Theta_j$. Unbiasedness implies

$$E_{\theta_0} V_i[\delta(X)] \leqq E_{\theta_0} V_j[\delta(X)],$$

$$E_{\theta_n} V_i[\delta(X)] \geqq E_{\theta_n} V_j[\delta(X)] \qquad \text{for } n = 1, 2, \cdots$$

and hence also

$$E_{\theta_0} V_i[\delta(X)] \geqq E_{\theta_0} V_j[\delta(X)].$$

It follows that

$$R_\delta(\theta_0) = E_{\theta_0} V_i[\delta(X)] = E_{\theta_0} V_j[\delta(X)] = \lim_{n\to\infty} R_\delta(\theta_n).$$

(iii) *Any restricted product of simple decision procedures is again simple.*

We can now prove the main result of this section, which provides a sufficient condition for unbiasedness of a product procedure to imply continuity of the risk functions of all of the component procedures.

(iv) Let $\mathcal{P}$ be a restricted product of a finite number of simple decision problems $\mathcal{P}_\gamma$, and suppose that the partitions Π_γ of the component problems into sets $\Theta_{i\gamma}$, $i \in I_\gamma$ satisfy the following condition.

> (*) Let $\Theta_{i\gamma_0}$, $\Theta_{j\gamma_0}$ be any two sets of Π_{γ_0} with common boundary points, let θ_0 be any such point, and assume without loss of generality that $\theta_0 \in \Theta_{i\gamma_0}$. Then there exists a sequence of points $\theta_n \in \Theta_{j\gamma_0}$ such that $\theta_n \to \theta_0$ and
>
> $$\theta_n \sim \theta_0(\Pi_\gamma) \quad \text{for all} \quad \gamma \neq \gamma_0,$$
>
> where $\theta \sim \theta'(\Pi)$ indicates that the two points lie in the same set of the partition.

Under these assumptions, if the risk function $R_\delta(\theta)$ of a product procedure δ is continuous, so are the risk functions $R_{\delta_\gamma}(\theta)$ of the components δ_γ of δ.

To see this, let γ_0 be any fixed value of γ, θ_0 any boundary point of Π_{γ_0}, and $\{\theta_n\}$ the sequence guaranteed by (*). Since $R_\delta(\theta) = \sum_\gamma R_{\delta_\gamma}(\theta)$ is continuous, we have

$$\sum_\gamma R_{\delta_\gamma}(\theta_n) \to \sum_\gamma R_{\delta_\gamma}(\theta_0).$$

Also, for each $\gamma \neq \gamma_0$ all of the points θ_n lie in the same set of the partition Π_γ with θ_0, so that by (i)

$$R_{\delta_\gamma}(\theta_n) \to R_{\delta_\gamma}(\theta_0) \quad \text{for all} \quad \gamma \neq \gamma_0.$$

It follows that $R(\delta_{\gamma_0},\theta_n) \to R(\delta_{\gamma_0},\theta_0)$, as was to be proved, where we have written $R(\delta, \theta)$ for $R_\delta(\theta)$.

It is convenient that (*) depends only on the partitions Π_γ, not on the values the loss functions W_γ take on over these partitions. For applications it is further important to note that (*) may be weakened slightly. Let $\Lambda_{i\gamma}$ be the set of common boundary points of $\Theta_{i\gamma}$ and $\Theta_{j\gamma}$ that belongs to $\Theta_{i\gamma}$. Then in order to ensure (iv) it is sufficient if (*) holds on a dense subset of each $\Lambda_{ij\gamma}$. This is an immediate consequence of (i).

Consider now any problem $\mathcal{P}$, which is a restricted product of a finite number of problems $\mathcal{P}_\pm(\theta_\gamma)$, and which satisfies (*). For the component problems continuity of the risk function is equivalent to similarity on the boundary at level $\alpha_\gamma = b_\gamma/(a_\gamma + b_\gamma)$. Hence a product procedure δ uniformly minimizes the risk among all unbiased procedures of $\mathcal{P}$ provided each component procedure δ_γ uniformly minimizes the risk among all procedures of $\mathcal{P}_\pm(\theta_\gamma)$ that are similar on the boundary at level α_γ. Since $\mathcal{P}_\pm(\theta_\gamma)$ is formally equivalent to the problem of testing $\theta \leqq \theta_\gamma$ or $\theta \geqq \theta_\gamma$, this is the case in particular if the possible distributions of the observable random variables constitute an exponential family, and the procedures δ_γ are the best unbiased tests of the hypotheses in question. Under

these conditions it is then only necessary to verify (*) in order to establish the desired optimum property for the resulting δ.

As an example consider problem (9.4), which is generated by $\mathcal{O}_{\pm}(\theta_0)$. Here $\mathcal{O}_{-}(\theta_0)$ induces the partition $\theta \leqq \theta_0$, $\theta > \theta_0$, with θ_0 as its only boundary point. Let θ_n be any sequence of points greater than and tending to θ_0. Then all the points θ_n and θ_0 lie in the set $\theta \geqq \theta_0$, and hence $\theta_n \sim \theta_0$ with respect to the partition induced by $\mathcal{O}_{+}(\theta_0)$, which completes the verification of (*). The argument is exactly the same for problem (9.6) and (10.1).

In the example of Sec. 10 leading to procedure (10.7), a typical partition is $\theta_2 \leqq \theta_1$, $\theta_2 > \theta_1$. Attention may be restricted to boundary points $\theta^{(0)}$ with coordinates

$$\theta^0_{i_1} < \cdots < \theta^{(0)}_{i_r} < \theta^{(0)}_1 = \theta^{(0)}_2 < \theta^{(0)}_{j_1} < \cdots < \theta^{(0)}_{j_{s-r-2}}.$$

For these, (*) is satisfied by the sequence of points $\theta^{(n)}$ with coordinates $\theta^{(n)}_i = \theta^{(0)}_i$ for $i \neq 2$, and $\theta^{(n)}_2$ between $\theta^{(0)}_1$ and $\theta^{(0)}_{j_1}$ and tending to $\theta^{(0)}_1$.

12. Consecutive decisions in a single experiment. The multiple decision problems treated in the previous sections were generated by the simultaneous consideration of a number of simpler component problems. We shall now suppose that these separate problems arise not in parallel but in sequence. A single sample is available for investigating a number of questions that are potentially of interest and are taken up one by one. Whether a given question is relevant, or which of a number of possible alternative formulations is appropriate at a certain stage, depends on the decisions reached up to that point.

(i) As an example suppose that independent variables $X_1, \cdots, X_n$ from a normal distribution $N(\xi, \sigma^2)$ are measurements on an experimental batch of a new product of quality ξ. The product is of no interest unless $\xi > \xi_0$, so that one will wish to test first of all the hypothesis $H_1: \xi \leqq \xi_0$. If the quality is found satisfactory, that is, if H_1 is rejected, it becomes necessary to investigate the variability of the product. One will then test $H_2: \sigma \geqq \sigma_0$, and in case this hypothesis is accepted one will try to reduce σ, for example by using less variable materials. The problem of testing H_2 arises here only in case H_1 is rejected.

(ii) Suppose that two treatments are being compared on a number of different categories of patients. Let the observed effect of treatment $i (i = 1, 2)$ on the kth patient of the jth category be distributed as $N(\xi_{ij}, \sigma^2)$ where

$$\xi_{ij} = \eta + \lambda_i + \mu_j + \nu_{ij} \left(\sum_i \lambda_i = \sum_j \beta_j = \sum_i \nu_{ij} = \sum_j \nu_{ij} = 0\right).$$

Here λ_i is the main effect of treatment i and ν_{ij} the interaction between the ith treatment and the jth category. One may believe in the possibility of the interactions being negligible and hence wish first to test the hypothesis $H_1: \nu_{ij} = 0$ for all i, j. In case H_1 is accepted the λ's are the objects of primary interest, and the problem becomes that of deciding whether $\lambda_2 - \lambda_1$ is $<$, $=$, or > 0, or to estimate this difference either by confidence intervals or by a point estimate. On the other hand, if H_1 is rejected one will be concerned less with the over-all effects of the treatments which is measured by the λ's than with the treatment differences $\xi_{2j} - \xi_{1j}$ for each category.

More generally, let there be given a first problem $\mathcal{P}'$, in which the possible decisions are $d_1', \cdots, d_m'$ and the loss function is $W'(\theta, d)$. If decision d_i is taken, a second problem $\mathcal{P}_i''$ arises with possible decisions $d_{ij}''(j = 1, \cdots, n_i)$ and loss function $W_i''(\theta, d)$. The combination of these two problems leads to a two-stage problem with decisions $d_{ij} = (d_i', d_{ij}'')$. One may of course continue in this manner and suppose that decision (d_i', d_{ij}'') gives rise to a further problem $\mathcal{P}_{ij}'''$ with decisions d_{ijk}'''. However, it is enough to treat the case of two levels since the discussion then extends immediately to the more general situation by induction.

In specifying a loss function we shall assume that even if a wrong decision is taken at the first step, so that the second problem is not the most appropriate one or perhaps need not have been considered at all, it is still desirable to do as well with respect to it as is possible. Thus in example (ii) above, if one has incorrectly decided that the interactions are negligible, one will in the estimation of $\theta_2 - \theta_1$ still wish to obtain as good an estimate as possible, and analogously if H_1 has wrongly been rejected. Whether the assumption holds in Example (i), that is, whether one would wish to control the variability of the new product after having mistakenly judged its quality to be satisfactory, appears to depend on the circumstances of the problem.

With this assumption, a natural loss function for the compound problem is

$$W(\theta, d_{ij}) = W'(\theta, d_i') + W_i''(\theta, d_{ij}''). \tag{12.1}$$

The possibility of not considering a second problem in case a certain decision d_i' is taken at the first step, is included in this formulation. One need then only take as problem $\mathcal{P}_i''$ the vacuous decision problem, that is, set $n_i = 1$ and $W_i''(\theta, d_{i1}'') = 0$. Suppose in particular, as was the case in Example (i), that a second problem occurs only for one of the decisions of the first stage, say d_1'. The possible decisions of the compound problem are then

$$d_{1j} = (d_1', d_{1j}''), \qquad j = 1, \cdots, n$$

and

$$d_2 = d_2', \cdots, d_m = d_m',$$

and the loss function is given by

$$\begin{aligned} W(\theta, d_i) &= W'(\theta, d_i'), && i = 2, \cdots, m; \\ W(\theta, d_{1j}) &= W'(\theta, d_1') + W''(\theta, d_{1j}''), && j = 1, \cdots, n. \end{aligned} \tag{12.2}$$

Returning now to the general case, suppose that there exist a satisfactory procedure δ' for $\mathcal{P}'$, which takes decision d_i' when $X \varepsilon D_i'$, and that the problems $\mathcal{P}_1'', \cdots, \mathcal{P}_m''$ are all different. It then seems natural to retain δ' as first step of the compound procedure, and to consider the problems at the second level relative to the circumstances in which they occur, namely conditionally given that $X \varepsilon D_1', \cdots, X \varepsilon D_m'$ respectively. Suppose further that for each $i = 1, \cdots, m$ there exists a satisfactory procedure δ_i'' for $\mathcal{P}_i''$ when the distribution of X is the conditional distribution given $X \varepsilon D_i'$. Such a procedure consists of a partition of the new sample space D_i' into regions D_{ij}'' in which the decisions

d''_{ij} are taken. Together, the sets $D''_{ij} (j = 1, \cdots, n_i ; i = 1, \cdots, m)$ form a partition of the original sample space and a solution of the compound problem, with decision $d_{ij} = (d'_i , d''_{ij})$ being taken when $X \varepsilon D''_{ij}$. One can of course again include in the formulation the possibility of ruling out some of the decisions d_{ij}, and the resulting compatibility questions can be treated exactly as before. However, this possibility seems to be less important in the present context and in order not to complicate the discussion unnecessarily we shall assume that no restrictions are imposed on the compound decision problem.

As an application consider the case that m and the n_i are equal to two, so that each of the component problems is one of hypothesis testing. Suppose that the hypotheses in question concern the parameters θ_i in an exponential family

$$dP_\theta(x) = C(\theta_1 , \cdots , \theta_r)e^{\sum \theta_i T_i(x)} \, d\nu(x).$$

There then exist uniformly most powerful unbiased tests of the hypotheses $\theta_i \leqq \theta_i^0$ and more generally of $\sum c_i\theta_i \leqq c_0$. (See, for example, [2].) Since after truncation on a fixed set D the family of distributions P_θ retains its property of forming an exponential family, such optimum unbiased tests will in particular also exist at the second stage after a preliminary test of significance has been performed as a first step. This will however not coincide with the standard optimum test for the corresponding problem without truncation.

We have assumed so far that the problems occurring at the second stage are all distinct. Suppose now instead that $\mathcal{P}''_1$ is appropriate when decisions $d'_1 , \cdots , d'_{r_1}$ are taken in the first problem, that $\mathcal{P}''_2$ corresponds to decisions $d'_{r_1+1} , \cdots , d'_{r_1+r_2}$, etc. One would then consider the problems $\mathcal{P}''_1$, $\mathcal{P}''_2$, $\cdots$ conditionally given that $X \varepsilon D'_1 + \cdots + D'_{r_1}$, $X \varepsilon D'_{r_1+1} + \cdots + D'_{r_1+r_2}$, $\cdots$, and otherwise proceed as before. If one is dealing in particular with the product of two decision problems so that all of the problems at the second level are the same, one considers this common problem $\mathcal{P}''$ given that $X \varepsilon D'_1 + \cdots + D'_m$, that is, unconditionally. The procedure therefore reduces to the product of the procedures for $\mathcal{P}'$ and $\mathcal{P}''$, so that the present theory agrees with that given earlier for products of decision problems.

Unfortunately the properties of the conditional procedures considered in the present section are not as satisfactory as of those discussed in the earlier parts of this paper. To be specific, let $\mathcal{P}'$ and $\mathcal{P}''_1 , \cdots , P''_m$ define a two-stage problem, in which the components $\mathcal{P}''_i$ of the second stage are distinct. The risk function of a procedure δ with components δ', $\delta''_1 , \cdots , \delta''_m$ is then

$$R_\delta(\theta) = R_{\delta'}(\theta) + \sum_{i=1}^m P_\theta(D'_i)R_{\delta''_i|\delta'}(\theta) \tag{12.3}$$

where the notation $R_{\delta''_i|\delta'}$ is used to indicate that this risk component is computed conditionally given $X \varepsilon D'_i$ and hence depends on δ' as well as on δ''_i.

It is clear from (12.3) that unbiasedness of δ' and $\delta''_1 , \cdots , \delta''_m$ implies that of δ. Also it is again true for most problems of interest that unbiasedness of δ implies either unbiasedness or at least similarity on the boundary for δ', $\delta''_1 , \cdots , \delta''_m$. However, the basic comparison of two procedures in terms of their

components is no longer simple. In particular,

$$R_{\delta'}(\theta) \leqq R_{\phi'}(\theta), \qquad R_{\delta_i''|\delta'}(\theta) \leqq R_{\phi_i''|\phi'}(\theta)$$

does not in general imply $R_\delta(\theta) \leqq R_\phi(\theta)$. If one wishes to minimize $R_\delta(\theta)$ then, given δ', one must select δ_i'' to minimize $R_{\delta_i''|\delta'}(\theta)$. But the best choice of δ' is not necessarily that which minimizes $R_{\delta'}(\theta)$ since the choice of δ' influences not only $R_{\delta'}(\theta)$ but also the second components of (12.3) and in particular the conditional risks $R_{\delta_i|\delta'}$. As a result of these complications it turns out that for the problem under consideration there usually does not exist among the unbiased procedures one that uniformly minimizes the risk. Of the procedures, the components of which have this optimum property we can only say that they are unbiased, and within the class of all unbiased procedures admissible.

13. Some examples of conditional procedures. Although we have found no satisfactory justification for the procedures discussed in the preceding section, they are rather natural from the Neyman-Pearson point of view, and we shall briefly illustrate them here with a few examples leaving a more detailed discussion and comparison with alternative procedures for a later paper.

(i) The problem mentioned at the beginning of Sec. 12 is concerned with testing, on the basis of a normal sample, the two hypotheses $H_1: \xi \leqq \xi_0$ and $H_2: \sigma \geqq \sigma_0$, where H_2 is assumed to be of interest only in case H_1 is rejected. If without loss of generality we put $\xi_0 = 0$, the best unbiased procedure for testing H_1 is Student's t-test with rejection region

$$\bar{X}/S \geqq C \tag{13.1}$$

and size $\alpha_1 = b_1/(a_1 + b_1)$. With this as first step, the condition of unbiasedness implies in the usual way that the rejection region R_2 of H_2 must satisfy

$$P_{\sigma_0}(R_2 \mid S \leqq \bar{x}/C \mid \bar{x}) \overset{(\bar{x})}{=} \alpha_2 . \tag{13.2}$$

Applying the fundamental lemma of Neyman and Pearson one sees that the uniformly most powerful unbiased conditional test of H_2 has a rejection region of the form

$$S^2 < f(\bar{X}). \tag{13.3}$$

Here the function f is defined by

$$\int_0^{f(u)} p_{\sigma_0}(z)\, dz = \alpha_2 \int_0^{u^2/C^2} p_{\sigma_0}(z)\, dz, \qquad u > 0, \tag{13.4}$$

where p_{σ_0} is the probability density of S^2 when $\sigma = \sigma_0$.

The resulting compound procedure then decides between the three possible conclusions

$d_1: \xi \leqq \xi_0$ —the new product is not of satisfactory quality,

$d_2: \xi > \xi_0, \sigma \geqq \sigma_0$ —the quality of the new product is satisfactory but its variability must be reduced,

$d_3: \xi > \xi_0, \sigma < \sigma_0$ —the new product is satisfactory with regard to both quality and variability,

as the sample falls into the corresponding one of the regions

$$D_1: \bar{X} \mid S \leqq C,$$
$$D_2: \bar{X} \mid S > C, \qquad S^2 \geqq f(\bar{X})$$
$$D_3: \bar{X} \mid S > C, \qquad S^2 < f(\bar{X}).$$

These decision regions are illustrated for the case $n = 10$, $\alpha_1 = \alpha_2 = .05$ in the figure.

(ii) As a somewhat more complex example consider two treatments which are being compared on a number of different categories of patients. Let the observed effect Y_{ijk} of treatment i ($i = 1, 2$) on the kth patient ($k = 1, \cdots, n$) in the jth category be distributed as $N(\xi_{ij}, \sigma^2)$, and let

$$\xi_{ij} = \eta + \lambda_i + \mu_j + \nu_{ij} \left(\sum_i \lambda_i = \sum_j \mu_j = \sum_i \nu_{ij} = \sum_j \nu_{ij} = 0\right)$$

where λ_i is the main effect of treatment i and ν_{ij} the interaction between the ith treatment and jth category. One may here wish to test first the hypothesis of no interaction

$$H_1: \nu_{ij} = 0 \qquad \text{for all } i, j.$$

If H_1 is accepted one will be interested in the difference of the λ's and wish either to test it or alternatively to estimate it by confidence intervals or point estimate. We shall here suppose that we then want to test

$$H_2: \lambda_2 - \lambda_1 \leqq 0.$$

On the other hand, if H_1 is rejected one will usually be concerned less with comparing the over-all effects of the two treatments, which is measured by the λ's, than with a comparison of the treatment effects ξ_{2j} and ξ_{1j} separately for each category j. In particular one may be interested to test the set of hypotheses

$$H_{3j}: \xi_{2j} - \xi_{1j} = (\lambda_2 + \nu_{2j}) - (\lambda_1 + \nu_{1j}) \leqq 0.$$

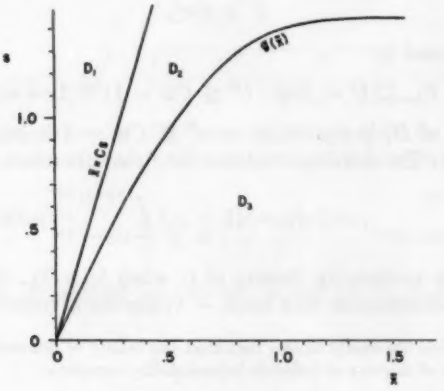

FIG. 1

Although there exists no uniformly most powerful unbiased test of H_1 it seems natural to start out with the standard test of this hypothesis which is uniformly most powerful invariant, and is given by the acceptance region

$$S_1^2/S_0^2 \leqq C \tag{13.5}$$

where

$$S_0^2 = \sum\sum\sum (Y_{ijk} - Y_{ij.})^2$$
$$S_1^2 = \sum\sum (Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2.$$

The hypothesis H_2 should then be considered conditionally, given $S_1^2/S_0^2 \leqq C$. A routine application of the theory of unbiased tests and similar regions[4] shows that a necessary condition for the rejection region R_2 to be unbiased is

$$P_{\lambda_2=\lambda_1}\{R_2 \mid s_1^2,\ s_0^2 + s_2^2,\ S_1^2/S_0^2 \leqq C\} = \alpha_2$$

for all values of s_1^2 and $s_0^2 + s_2^2$ where

$$S_2^2 = \textstyle\sum_{i=1}^{2} (Y_{i..} - Y_{...})^2 = \frac{1}{2}(Y_{2..} - Y_{1..})^2.$$

If one puts

$$U = (Y_{2..} - Y_{1..})/S_0, \qquad V = S_0^2 + S_2^2, \qquad W = (S_0^2 + S_2^2)/S_1^2,$$

it follows from the fundamental lemma that the uniformly most powerful (conditional) test of H_2 has the rejection region

$$U \geqq k(V, W)$$

where k is determined by

$$P_{\lambda_1=\lambda_2}\{U \geqq k(V, W) \mid V = v,\ W = w \quad \text{and} \quad U^2 \leqq Cw - 1\}^{(v,w)}_{\alpha_2}.$$

Now when $\lambda_1 = \lambda_2$, the variable U is independent of V so that k depends only on w, $k(v, w) = f(w)$ say. The rejection region then becomes

$$U \geqq f(w), \tag{13.6}$$

where f is determined by

$$P_{\lambda_1=\lambda_2}\{U < f(w) \mid U^2 \leqq Cw - 1\}^{(w)}_{=}(1 - \alpha_2).$$

Since acceptance of H_1 is equivalent to $u^2 \leqq Cw - 1$ it implies in particular that $0 \leqq Cw - 1$. The defining condition for f may therefore be written as

$$\int_{-(Cw-1)^{1/2}}^{f(w)} p_U(u)\,du = (1 - \alpha_2) \int_{-(Cw-1)^{1/2}}^{(Cw-1)^{1/2}} p_U(u)\,du \tag{13.7}$$

where $p_U(u)$ is the probability density of U when $\lambda_1 = \lambda_2$, that is, essentially, the density of a t-distribution with $2m(n - 1)$ degrees of freedom.

[4] The proof requires the easily shown fact that the family of noncentral χ^2-distributions with a fixed number of degrees of freedom is boundedly complete.

Let us now consider in a similar manner the hypothesis

$$H_{31}: (\lambda_2 - \lambda_1) + (\nu_{21} - \nu_{11}) \leqq 0,$$

conditionally given that $S_1^2/S_0^2 > C$. Let $\hat{\lambda}_i$ and $\hat{\nu}_{ij}$ be the least squares estimates of the corresponding parameters so that

$$S_1^2 = \sum_{i=1}^{2} \sum_{j=1}^{m} \hat{\nu}_{ij}^2$$

and let

$$\begin{aligned} Z_1 &= \tfrac{1}{2}(\hat{\lambda}_2 - \hat{\lambda}_1 - \hat{\nu}_{21} + \hat{\nu}_{11}) \\ Z_2 &= \tfrac{1}{2}(\hat{\lambda}_2 - \hat{\lambda}_1 + \hat{\nu}_{21} - \hat{\nu}_{11}) \\ S_1'^2 &= S_1^2 - (\hat{\nu}_{21}^2 - \hat{\nu}_{11})^2 = S_1^2 - (Z_2 - Z_1)^2. \end{aligned}$$

If $\zeta_i = E(Z_i)$, the hypothesis becomes $\zeta_2 \leqq 0$, and unbiasedness of the conditional test of H_{31} with rejection region R_{31} implies

$$P_{\zeta_2=0}\{R_{31} \mid z_1, s_0^2 + z_2^2, s_1'^2 \quad \text{and} \quad S_1^2 / S_0^2 > C\} \equiv \alpha_{31}.$$

The condition $S_1^2 > CS_0^2$ is equivalent to $(Z_2 - Z_1)^2 > CS_0^2 - S_1'^2$, which may be rewritten as

$$\left(Z_2 - \frac{Z_1}{1+C}\right)^2 + \frac{CZ_1^2}{(1+C)^2} > \frac{C}{1+C} T^2 - S_1'^2,$$

where $T^2 = S_0^2 + Z_2^2$. This is satisfied for all values of Z_2 if

$$\frac{C}{1+C} T^2 - S_1'^2 - \frac{CZ_1^2}{(1+C)^2} \leqq 0 \tag{13.8}$$

and otherwise for the values of Z_2, for which either

$$\frac{Z_2}{T} > \frac{Z_1}{(1+C)T} + \sqrt{\frac{C}{1+C} - \frac{S_1'^2}{T^2} - \frac{C}{(1+C)^2}\frac{Z_1^2}{T^2}} = K_2\left(\frac{Z_1}{T}, \frac{S_1'}{T}\right)$$

or

$$\frac{Z_2}{T} < \frac{Z_1}{(1+C)T} - \sqrt{\frac{C}{1+C} - \frac{S_1'^2}{T^2} - \frac{C}{(1+C)^2}\frac{Z_1^2}{T^2}} = K_1\left(\frac{Z_1}{T}, \frac{S_1'}{T}\right).$$

Since Z_2/T is independent of Z_1/T and S_1'/T when $\zeta_2 = 0$, the uniformly most powerful unbiased test of H_{31} given $S_1^2/S_0^2 > C$ is then given by the rejection region

$$\frac{Z_2}{T} \geqq K\left(\frac{Z_1}{T}, \frac{S_1'}{T}\right)$$

with the function K defined as follows. When $(Z_1/t, S_1'/t)$ satisfies (13.8), we have

$$K\left(\frac{z_1}{t}, \frac{s_1'}{t}\right) = K$$

where

$$\int_K^1 p_R(r)\,dr = \alpha_{31},$$

$p_R(r)$ being the probability density of

$$R = \frac{Z_2}{T} = \frac{\frac{1}{2}(\hat{\lambda}_2 - \hat{\lambda}_1 - \hat{\nu}_{21} + \hat{\nu}_{11})}{\sqrt{S_0^2 + \frac{1}{4}(\hat{\lambda}_2 - \hat{\lambda}_1 + \hat{\nu}_{21} - \hat{\nu}_{11})^2}}$$

when $\zeta_2 = 0$. For all other values of $(z_1/t, s_1'/t)$, $K(z_1/t, s_1'/t)$ is given by

$$\int_{K((z_1/t),(s_1'/t))}^1 p_R(r)\,dr = \alpha_{31}\left[\int_{-1}^{K_1((z_1/t),(s_1'/t))} p_R(r)\,dr + \int_{K_2((z_2/t),(s_1'/t))}^1 p_R(r)\,dr\right].$$

(iii) As an example involving more than two stages let us consider the determination of the degree of a regression polynomial. Let $Y_1, \cdots, Y_n$ be independently normally distributed with constant variance σ^2 and means

$$\eta_i = E(Y_i) = e_s + e_{s-1}x_i + \cdots + e_0 x_i^s \qquad (i = 1, \cdots, n).$$

We shall assume that a polynomial of degree s will in any case be adequate for our purposes, and wish to determine the smallest degree $r \leqq s$ that would also be satisfactory. It is convenient for this purpose to express the regression polynomial in terms of the orthogonal polynomials P_i defined by

$$\sum_{k=1}^n P_i(x_k)P_j(x_k) = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j. \end{cases}$$

Writing

$$\eta_i = c_s + c_{s-1}P_1(x_i) + \cdots + c_0 P_s(x_i),$$

we test successively the hypotheses $H_0: c_0 = 0$, $H_1: c_1 = 0, \cdots$, continuing as long as no rejection occurs.

The problem can be stated in the following canonical form: $X_1, \cdots, X_s$ and S_0^2 are independent variables, with X_i being distributed as $N(\xi_i, \sigma^2)$ and with S_0^2/σ^2 having a χ^2-distribution with $n_0 = n - s$ degrees of freedom. One wishes to test consecutively the hypotheses $H_i: \xi_i = 0$, $(i = 1, \cdots, s)$ continuing as long as no rejection occurs. Slightly more generally one might have variables $X_{ij}: N(\xi_{ij}, \sigma^2)$, $(j = 1, \cdots, n_i\,;\, i = 1, \cdots, s)$ and S_0^2 and consider consecutively the hypotheses $H_i: \xi_{ij} = 0$ for $j = 1, \cdots, n_i$. Invariance reduces the problem to the statistics S_0^2 and $S_i^2 = \sum_{j=1}^{n_i} X_{ij}^2$, which are independent and where S_i^2/σ^2 has a χ^2 distribution with n_i degrees of freedom when H_i is true.

Let

$$U_i = \frac{S_i^2}{S_0^2}, \qquad V_i = \frac{S_i^2}{S_{i-1}^2}, \qquad W_i = V_i\left(1 + \frac{1}{U_i}\right) = \frac{S_i^2 + S_0^2}{S_{i-1}^2}.$$

We shall now show inductively that the best unbiased conditional test of H_i, given the acceptance of $H_1, \cdots, H_{i-1}$, has an acceptance region of the form

$$U_i \leqq f_i(W_i, V_{i-1}, \cdots, V_2). \tag{13.9}$$

The functions f_i are defined by

$$\begin{cases} f_1(w_1) = C, \\ \displaystyle\int_0^{f_i(w_i, v_{i-1}, \cdots, v_2)} p_i(r)\, dr = (1 - \alpha_i) \int_0^{w_i h_{i-1}(v_{i-1}, \cdots, v_2) - 1} p_i(r)\, dr \quad (i \geqq 2) \end{cases} \tag{13.10}$$

where p_i is the probability density of S_i^2/S_0^2 when H_i is true, h_i is defined by

$$\begin{cases} h_1 = C, \\ h_i(v_i, \cdots, v_2) = \min[v_i h_{i-1}(v_{i-1}, \cdots, v_2), g_i(v_i, \cdots, v_2)] \quad (i \geqq 2) \end{cases} \tag{13.11}$$

and g_i is a function taking on values in the space of u_i and defined by

$$u = g_i(v_i, \cdots, v_2) \leftrightarrow v_i = \frac{f_i^{-1}(u, v_{i-1}, \cdots, v_2)}{1 + \dfrac{1}{u}} \qquad (i \geqq 2) \tag{13.12}$$

where f_i^{-1} denotes the inverse of f_i considered for fixed values of $v_{i-1}, \cdots, v_2$ as a function of u. It is seen that (13.10), (13.11) and (13.12) define f_i in terms of f_{i-1}.

The best unbiased test of H_1 has the acceptance region $u_1 \leqq C$ and hence satisfies (13.9). Suppose now that the acceptance regions $A_1, \cdots, A_{i-1}$ have been shown to be given by (13.9). To prove (13.9) for A_i we need the fact that

$$A_1 \cap \cdots \cap A_{i-1} = \{(u, v) : u_{i-1} \leqq h_{i-1}(v_{i-1}, \cdots, v_2)\} \tag{13.13}$$

and since this is true for $i = 2$, we may accept (13.13) as part of our induction hypothesis. The condition of unbiasedness implies that A_i should satisfy

$$P_{H_i}\{A_i \mid s_0^2 + s_i^2, s_1^2, \cdots, s_{i-1}^2 \quad \text{and} \quad U_{i-1} \leqq h_{i-1}(V_{i-1}, \cdots, V_2)\} = 1 - \alpha_i.$$

Now $U_{i-1} \leqq h_{i-1}$ is equivalent to

$$U_i \leqq W_i h_{i-1}(V_{i-1}, \cdots, V_2) - 1$$

so that the best unbiased acceptance region for H_i is

$$U_i \leqq K_i(W_i, S_1^2, \cdots, S_{i-1}^2)$$

where

$$P_{H_i}\{U_i \leqq K_i \mid s_0^2 + s_i^2, s_1^2, \cdots, s_{i-1}^2 \quad \text{and} \quad U_i \leqq W_i h_{i-1}(V_{i-1}, \cdots, V_2) - 1\} = 1 - \alpha_i.$$

Since U_i is independent of $W_i, V_{i-1}, \cdots, V_2$ when H_i is true, it is seen that K_i depends only on $w_i, v_{i-1}, \cdots, v_2$ and that A_i is given by (13.9).

To complete the proof, it only remains to verify (13.13) for the set

$A_1 \cap \cdots \cap A_i$, which is the intersection of

$$A_1 \cap \cdots \cap A_{i-1}\colon u_{i-1} \leqq h_{i-1}(v_{i-1}, \cdots, v_2)$$

and

$$A_i\colon u_i \leqq f_i(w_i, v_{i-1}, \cdots, v_2).$$

The inequality describing A_i is equivalent to

$$f_i^{-1}(u_i, v_{i-1}, \cdots, v_2) \leqq w_i = v_i\left(1 + \frac{1}{u_i}\right)$$

and hence to

$$A_i\colon u_i \leqq g_i(v_i, \cdots, v_2)$$

by (13.12). Since the inequality describing $A_1 \cap \cdots \cap A_{i-1}$ is equivalent to

$$A_1 \cap \cdots \cap A_{i-1}\colon u_i \leqq v_i h_{i-1}(v_{i-1}, \cdots, v_2),$$

the intersection $A_1 \cap \cdots \cap A_i$ is given by (13.13), as was to be proved.

A closely related problem arises in the study of components of variance when one is dealing with a hierarchical classification. Here the problem reduces to independent statistics S_1^2, S_2^2, $\cdots$ where S_i^2/σ_i^2 has a χ^2-distribution with n_i degrees of freedom. It follows from the underlying model that $\sigma_1 \leqq \sigma_2 \leqq \cdots$, and one is interested in testing successively the hypotheses $H_1\colon \sigma_2 = \sigma_1$, $H_2\colon \sigma_3 = \sigma_2, \cdots$, continuing only if all the previous hypotheses were accepted. If $H_1, \cdots, H_i$ are true, the distribution of $S_1^2, \cdots, S_i^2$ is the same as in the preceding problem, and it is easily seen that exactly the same procedure is applicable to the present situation.

14. Minimizing the maximum risk. For problems of the kind discussed in the preceding section unbiasedness is closely related to the minimax property. We begin by considering the two-decision problem of testing a hypothesis $H\colon \theta \in \omega$. Let d_0 and d_1 denote the decisions of accepting and rejecting H, and let the losses of false rejection and acceptance be a and b respectively. Then we have

(i) A necessary condition for a procedure δ to be minimax is that it be unbiased.

(ii) This condition is also sufficient if the probability $P_\theta(A)$ of any set A is continuous in θ and if the common boundary of ω and ω^{-1} is nonempty.

To see (i) note that unbiasedness of a procedure δ_0 implies

$$P_\theta(d_1) \leqq \frac{b}{a+b} \text{ for } \theta \in \omega, \qquad P_\theta(d_0) \leqq \frac{a}{a+b} \text{ for } \theta \in \omega^{-1}. \tag{14.1}$$

Since the risk function of any procedure δ is given by

$$R_\delta(\theta) = \begin{cases} aP_\theta(d_1) & \text{for} \quad \theta \in \omega \\ bP_\theta(d_0) & \text{for} \quad \theta \in \omega^{-1}, \end{cases} \tag{14.2}$$

δ_0 satisfies

$$\sup R_{\delta_0}(\theta) \leqq \frac{ab}{a+b}. \tag{14.3}$$

Any minimax procedure δ_1 must then also satisfy (14.3) and hence by (14.2) must also be unbiased. Suppose next that the assumptions of (ii) hold, let θ_0 be a common boundary point of ω and ω^{-1}, and let δ_0 be unbiased. Then (14.1) and $P_\theta(d_0) = 1 - P_\theta(d_1)$ show that

$$P_{\theta_0}(d_1) = \frac{b}{a+b} \text{ and hence } R_{\delta_0}(\theta_0) = ab/(a+b).$$

Thus for any unbiased procedure δ_0 we have

$$\sup R_{\delta_0}(\theta) = \frac{ab}{a+b}. \tag{14.4}$$

By (i), this relation holds in particular for any minimax procedure, so that $ab/(a+b)$ is the minimax risk and (ii) follows from (14.4).

We shall now extend (ii) to the case of a number of hypotheses $H_i: \xi_i(\theta) \in \omega_i$ $(i = 1, \cdots, s)$ to be tested in sequence, where each hypothsis H_i is tested only if a particular prescribed chain of decisions has been reached on $H_1, \cdots, H_{i-1}$. For these problems we have

THEOREM 3. *Every unbiased procedure minimizes the maximum risk provided* (i) $P_\theta(A)$ *is a continuous function of* θ *for each* A, (ii) *the* 2^s *intersections* $\bigcap_{i=1}^{s} \omega_i^{x_i}$, *(each* $x_i = \pm 1$*), are all nonempty,* (iii) *there exists at least one parameter point* θ_0 *which lies on the common boundary of* ω_i *and* ω_i^{-1} *for all* i, (iv) *the losses* a_i *and* b_i *for falsely rejecting and accepting* H_i *satisfy condition* (14.10) *below.*

PROOF. Let us write a_i^{-1} for b_i, let d_i and d_i^{-1} denote the decisions of rejecting and accepting H_i, and let $y_i = +1$ if H_{i+1} is considered in case H_i is rejected, and $y_i = -1$ in the contrary case. If $\theta \in \bigcap_{i=1}^{s} \omega_i^{x_i}$, the risk of a procedure δ is then given by

$$\begin{aligned} R_\delta(\theta) &= a_1^{x_1} P(d_1^{x_1}) + P(d_1^{y_1})\{a_2^{x_2} P(d_2^{x_2} \mid d_1^{y_1}) + P(d_2^{y_2} \mid d_1^{y_1}) \\ &\qquad [a_3^{x_3} P(d_3^{x_3} \mid d_1^{y_1} d_2^{y_2} + \cdots]\} \\ &= a_1^{x_1} P(d_1^{x_1}) + a_2^{x_2} P(d_1^{y_1}) P(d_2^{x_2} \mid d_1^{y_1}) \\ &\qquad + a_3^{x_3} P(d_1^{y_1}) P(d_2^{y_2} \mid d_1^{y_1}) P(d_3^{x_3} \mid d_1^{y_1} d_2^{y_2}) + \cdots. \end{aligned}$$

By comparing each of these expressions for which $x_i = 1$ with the corresponding one for which $x_i = -1$ but all the other x's are the same, it is seen that unbiasedness implies and hence is equivalent to the conditions

$$P_\theta(d_i \mid d_1^{y_1} \cdots d_{i-1}^{y_{i-1}}) \leqq \frac{b_i}{a_i + b_i} \text{ for } \theta \in \omega_i$$

$$P_\theta(d_i^{-1} \mid d_1^{y_1} \cdots d_{i-1}^{y_{i-1}}) \leqq \frac{a_i}{a_i + b_i} \text{ for } \theta \in \omega_i^{-1}$$

for $i = 1, \cdots, s$. Putting $r_i = a_i b_i / (a_i + b_i)$ we see that the risk function of any unbiased procedure satisfies

$$\sup R_\delta(\theta) \leqq r_1 + \frac{a_1^{y_1}}{a_1 + b_1} r_2 + \frac{a_1^{y_1}}{a_1 + b_1}\frac{a_2^{y_2}}{a_2 + b_2} r_3 + \cdots + \frac{a_1^{y_1}}{a_1 + b_1} \cdots \frac{a_{s-1}^{y_{s-1}}}{a_{s-1} + b_{s-1}} r_s = r^*.$$

It now remains to show that r^* is the minimax risk.

Consider any procedure δ, and let

$$\alpha_i = P_{\theta_0}(d_i \mid d_1^{y_1} \cdots d_{i-1}^{y_{i-1}}).$$

Then it follows from assumptions (i), (ii) and (iii) that

$$\sup R_\delta(\theta) \geqq \max_{x_1, \cdots, x_s} [(a_1\alpha_1)^{x_1} + \alpha_1^{y_1}(a_2\alpha_2)^{x_2} + \cdots + \alpha_1^{y_1} \cdots \alpha_{s-1}^{y_{s-1}}(a_s\alpha_s)^{x_s}] \tag{14.6}$$

since the 2^s expressions in brackets are the possible values of $\lim R_\delta(\theta)$ as θ tends to θ_0. Let us now minimize the right-hand side of (14.6). Given any $\alpha_1, \cdots, \alpha_{s-1}$, $x_1, \cdots, x_{s-1}$, the maximum of the pair of expressions

$$(a_1\alpha_1)^{x_1} + \cdots + \alpha_1^{y_1} \cdots \alpha_{s-2}^{y_{s-2}}(a_{s-1}\alpha_{s-1})^{x_{s-1}} + \alpha_1^{y_1} \cdots \alpha_{s-1}^{y_{s-1}}(a_s\alpha_s),$$

$$(a_1\alpha_1)^{x_1} + \cdots + \alpha_1^{y_1} \cdots \alpha_{s-2}^{y_{s-2}}(a_{s-1}\alpha_{s-1})^{x_{s-1}} + \alpha_1^{y_1} \cdots \alpha_{s-1}^{y_{s-1}}(a_s\alpha_s)^{-1},$$

is clearly minimized by minimizing

$$\max[a_s\alpha_s, a_s^{-1}(1 - \alpha_s)].$$

This gives $\alpha_s = \alpha_s^*$ where we put

$$\alpha_i^* = \frac{a_i^{-1}}{a_i + a_i^{-1}} = \frac{b_i}{a_i + b_i}, \tag{14.7}$$

so that

$$a_i\alpha_i^* = (a_i\alpha_i^*)^{-1}. \tag{14.8}$$

We can now proceed inductively. Suppose that it has already been shown for any fixed $\alpha_1, \cdots, \alpha_i$ that the right-hand side of (14.6) is minimized by $\alpha_j = \alpha_j^*$ for $j = i + 1, \cdots, s$. Consider now the minimization of the maximum of the quantities

$$(a_1\alpha_1)^{x_1} + \cdots + \alpha_1^{y_1} \cdots \alpha_{i-1}^{y_{i-1}}(a_i\alpha_i) + \alpha_1^{y_1} \cdots \alpha_i^{y_i}[(a_{i+1}\alpha_{i+1}^*)^{x_{i+1}} + (\alpha_{i+1}^*)^{y_{i+1}}(a_{i+1}\alpha_{i+2}^*)^{x_{i+2}} + \cdots],$$

$$(a\alpha_1)^{x_1} + \cdots + a_1^{y_1} \cdots \alpha_{i-1}^{y_{i-1}}(a_i\alpha_i)^{-1} + \alpha_1^{y_1} \cdots \alpha_i^{y_i} [(a_{i+1}\alpha_{i+1}^*)^{x_{i+1}} + (\alpha_{i+1}^*)^{y_{i+1}}(a_{i+2}\alpha_{i+2}^*)^{y_{i+2}} + \cdots],$$

where the expression in brackets by (14.8) is independent of the x's and hence equal to

$$(14.9)\quad k_i = a_{i+1}\alpha_{i+1}^* + (\alpha_{i+1}^*)^{y_{i+1}}a_{i+2}\alpha_{i+2}^* + \cdots + (\alpha_{i+1}^*)^{y_{i+1}}\cdots(\alpha_{s-1}^*)^{y_{i-1}}a_s\alpha_s^*.$$

Eliminating the common terms at the beginning, and the common factor $\alpha_1^{y_1}\cdots\alpha_{i-1}^{y_{i-1}}$ of the terms involving α_i, we see that it is enough to minimize

$$\max[a_i\alpha_i + k_i\alpha_i^{y_i}, b_i\alpha_i^{-1} + k_i\alpha_i^{y_i}].$$

This is achieved by equating the two quantities, which gives $\alpha_i = \alpha_i^*$, provided the coefficient of α_i in $b_i + (k_i - b_i)\alpha_i$ is <0 in case $y_i = 1$, or that the coefficient of α_i in $a_i\alpha_i + k_i(1 - \alpha_i)$ is >0 in case $y_i = -1$. Thus the right-hand side of (14.6) is minimized by putting $\alpha_i = \alpha_i^*$ for all i, provided

$$(14.10)\quad a_{i+1}\alpha_{i+1}^* + (\alpha_{i+1}^*)^{y_{i+1}}a_{i+2}\alpha_{i+2}^* + \cdots + (\alpha_{i+1}^*)^{y_{i+1}}\cdots(\alpha_{s-1}^*)^{y_{s-1}}a_s\alpha_s^* < a_i^{-y_i}.$$

Putting $\alpha_i = \alpha_i^*$ in the right-hand side of (14.6) one then sees that the 2^r quantities in brackets become equal, and that their common value is r^*. Thus, for any procedure δ, we have

$$\sup R_\delta(\theta) \geqq r^*$$

and the desired result now follows from comparison with (14.5).

COROLLARY. *The conclusion of Theorem* 3 *holds if* (14.10) *is replaced by either*

$$(14.11)\quad y_i = 1 \text{ for all } i \quad \text{and} \quad b_1 \geqq b_2 \geqq \cdots \geqq b_s$$

or

$$(14.12)\quad y_i = -1 \text{ for all } i \quad \text{and} \quad a_1 \geqq a_2 \geqq \cdots \geqq a_s.$$

PROOF. It is necessary only to show that each of the conditions (14.11) and (14.12) implies (14.10). If all the y's are $+1$ the left-hand side of (14.10) may be rewritten as

$$b_{i+1}(\alpha_{i+1}^*)^{-1} + b_{i+2}(\alpha_{i+2}^*)^{-1}\alpha_{i+1}^* + \cdots + b_s(\alpha_s^*)^{-1}\alpha_{i+1}^*\cdots\alpha_{s-1}^*.$$

If (14.11) holds, this is

$$\begin{aligned}&\leqq b_{i+1}[(\alpha_{i+1}^*)^{-1} + \alpha_{i+1}^*(\alpha_{i+2}^*)^{-1} + \cdots + \alpha_{i+1}^*\cdots\alpha_{s-1}^*(\alpha_s^*)^{-1}]\\ &= b_{i+1}[1 - \alpha_{i+1}^*\alpha_{i+2}^*\cdots\alpha_s^*] < b_{i+1} \leqq b_i.\end{aligned}$$

Similarly, if (14.12) holds, the left-hand side of (14.10) is

$$\begin{aligned}&\leqq a_{i+1}[\alpha_{i+1}^* + (\alpha_{i+1}^*)^{-1}\alpha_{i+2}^* + \cdots + (\alpha_{i+1}^*\cdots\alpha_{s-1}^*)^{-1}\alpha_s^*]\\ &= a_{i+1}[1 - (\alpha_{i+1}^*\cdots\alpha_s^*)^{-1}] < a_i.\end{aligned}$$

If the assumptions of Theorem 3 are not satisfied it is sometimes possible to prove a slightly weaker result. We shall illustrate this with the simplest case of Example (ii) of Sec. 12. Here one is concerned with four means ξ_{ij} $(i, j = 1, 2)$ given by

$$\begin{aligned}\xi_{11} &= \lambda + \mu + \nu, \qquad \xi_{12} = \lambda - \mu - \nu\\ \xi_{21} &= -\lambda + \mu - \nu, \qquad \xi_{22} = -\lambda - \mu - \nu.\end{aligned}$$

The first hypothesis tested is $H_0: \nu = 0$. In case of acceptance this is followed by $H_1: \lambda \leqq 0$, while in case of rejection one is interested in the two hypotheses $H_2: \xi_{11} \leqq \xi_{21}$ which is equivalent to $\lambda + \nu \leqq 0$ or $H_3: \xi_{12} \leqq \xi_{22}$ which is equivalent to $\lambda - \nu \leqq 0$. The feature which complicates this problem is that when H_0 is true the three remaining hypotheses must either all be true or all be false. Thus the condition corresponding to (ii) of Theorem 3 is not satisfied.

Let us denote the decision of rejecting or accepting H_i by d_i and $\bar{d}_i^{-1}$ as before, and consider the class $\mathfrak{C}_0$ of procedures satisfying

$$
\begin{cases}
P(d_0) \leqq \dfrac{b_0}{a_0 + b_0} \quad \text{if} \quad \nu = 0, \qquad P(\bar{d}_0^{-1}) \leqq \dfrac{a_0}{a_0 + b_0} \quad \text{if} \quad \nu \neq 0 \\
P(d_1 \mid \bar{d}_0^{-1}) \leqq \dfrac{b_1}{a_1 + b_1} \quad \text{if} \quad \lambda + \nu \leqq 0, \\
\qquad\qquad P(\bar{d}_1^{-1} \mid \bar{d}_0^{-1}) \leqq \dfrac{a_1}{a_1 + b_1} \quad \text{if} \quad \lambda + \nu > 0 \\
P(d_i \mid d_0) \leqq \dfrac{b_i}{a_i + b_i} \quad \text{if} \quad H_i \text{ is true}, \\
\qquad\qquad P(\bar{d}_i^{-1} \mid d_0) \leqq \dfrac{a_i}{a_i + b_i} \quad \text{if} \quad H_i \text{ is false} \quad (i = 2, 3),
\end{cases}
\tag{14.13}
$$

We shall then prove that any element of $\mathfrak{C}_0$, which is a subclass of the class of unbiased procedures, is minimax, under the additional assumption that

$$a_i = a \quad \text{and} \quad b_i = b \text{ for all } i.$$

For any procedure δ, consider the error probabilities $\alpha_0 = P(d_0)$, $\alpha_1 = P(d_1 \mid \bar{d}_0^{-1})$ $\alpha_i = P(d_i \mid d_0)$ $(i = 2, 3)$, evaluated for $\lambda = \nu = 0$ and some fixed values μ_0 and σ_0 of μ and σ. Then the possible value of $\lim R_\delta(\theta)$ as $\theta = (\lambda, \mu, \nu, \sigma)$ tends to $\theta_0 = (0, \mu_0, 0, \sigma_0)$ are

$$
\begin{aligned}
&a_1(1 - \alpha_0)\alpha_1 + \alpha_0[a_0 + a_2\alpha_2 + a_3\alpha_3] && \text{for } \theta \in \omega_0\omega_1\omega_2\omega_3 \\
&b_1(1 - \alpha_0)(1 - \alpha_1) + \alpha_0[a_0 + b_2(1 - \alpha_2) + b_3(1 - \alpha_3)] \\
&\qquad\qquad \text{for } \theta \in \omega_0\omega_1^{-1}\omega_2\omega_3^{-1} \\
&(1 - \alpha_0)[b_0 + a_1\alpha_1] + \alpha_0[a_2\alpha_2 + a_3\alpha_3] && \text{for } \theta \in \omega_0^{-1}\omega_1\omega_2\omega_3 \\
&(1 - \alpha_0)[b_0 + a_1\alpha_1] + \alpha_0[a_2\alpha_2 + b_3(1 - \alpha_3)] && \text{for } \theta \in \omega_0^{-1}\omega_1\omega_2\omega_3^{-1} \\
&(1 - \alpha_0)[b_0 + a_1\alpha_1] + \alpha_0[b_2(1 - \alpha_2) + a_3\alpha_3] && \text{for } \theta \in \omega_0^{-1}\omega_1\omega_2^{-1}\omega_3 \\
&(1 - \alpha_0)[b_0 + a_1\alpha_1] + \alpha_0[b_2(1 - \alpha_2) + b_3(1 - \alpha_3)] \\
&\qquad\qquad \text{for } \theta \in \omega_0^{-1}\omega_1\omega_2^{-1}\omega_3^{-1} \\
&(1 - \alpha_0)[b_0 + b_1(1 - \alpha_1)] + \alpha_0[a_2\alpha_2 + a_3\alpha_3] && \text{for } \theta \in \omega_0^{-1}\omega_1^{-1}\omega_2\omega_3
\end{aligned}
\tag{14.14}
$$

$$(1 - \alpha_0)[b_0 + b_1(1 - \alpha_1)] + \alpha_0[a_2\alpha_2 + b_3(1 - \alpha_3)]$$

for $\theta \in \omega_0^{-1}\omega_1^{-1}\omega_2\omega_3^{-1}$

$$(1 - \alpha_0)[b_0 + b_1(1 - \alpha_1)] + \alpha_0[b_2(1 - \alpha_2) + a_3\alpha_3]$$

for $\theta \in \omega_0^{-1}\omega_1^{-1}\omega_2^{-1}\omega_3$

$$(1 - \alpha_0)[b_0 + b_1(1 - \alpha_1)] + \alpha_0[b_2(1 - \alpha_2) + b_3(1 - \alpha_3)]$$

for $\theta \in \omega_0^{-1}\omega_1^{-1}\omega_2^{-1}\omega_3^{-1}$.

With $u = \alpha_0$, $v = (1 - \alpha_0)\alpha_1$, $w = \alpha_0\alpha_2$, $z = \alpha_0\alpha_3$ and $a_i = a$, $b_i = b$ these quantities become

$$\begin{aligned}
&a[u + v + w + z] \\
&(a + b)u - bv - bw - bz + b \\
&-bu + av + aw + az + b \\
&av + aw - bz + b \\
&av - bw + az + b \\
&bu + av - bw - bz + b \\
&-2bu - bv + aw + az + 2b \\
&-bu - bv + aw - bz + 2b \\
&-bu - bv - bw + az + 2b \\
&-2bu - bv - bw - bz + 2b.
\end{aligned} \tag{14.15}$$

From the first form of these 10 quantities it is seen that they all become equal for $\alpha_i = \alpha_i^* = b_i / (a_i + b_i)$. Let the corresponding values of (u, v, w, z) be (u_0, v_0, w_0, z_0). We shall now show that if we change from (u_0, v_0, w_0, z_0) to some other point (u_1, v_1, w_1, z_1) at least one of the 10 quantities will be increased so that (u_0, v_0, w_0, z_0) and hence $\alpha_i = \alpha_i^*$ minimizes their maximum. If in (14.15) all possible sign combinations were occurring, this would be obvious. For then in at least one row all of the four increments $\pm(u_1 - u_0)$, $\pm(v_1 - v_0)$, $\pm(w_1 - w_0)$, $\pm(z_1 - z_0)$ would be positive. But of the 16 possible combinations only 12 occur (with rows 4 and 5 each counting for two combinations, the missing combinations being $- + - -$, $+ - + +$, $+ - + -$, $+ - - +$.

As an example let us consider the first possibility. Suppose that $u_1 = u_0 - \xi$, $v_1 = v_0 + \eta$, $w_1 = w_0 - \Delta$, $z_1 = z_0 - \zeta$, with ξ, η, Δ, ζ all nonnegative. The total change in the first and tenth rows of (14.15) will be

$$a(-\xi + \eta - \Delta - \zeta) \quad \text{and} \quad b(2\xi - \eta + \Delta + \zeta).$$

Since both of these are to be $\leqq 0$, we have

$$2\xi + \Delta + \zeta \leqq \eta \leqq \xi + \Delta + \zeta$$

and hence $\xi = 0$. But then the change in the sixth row will be positive unless also $\eta = \Delta = \zeta = 0$. The other three possibilities can be ruled out in a similar manner so that $\alpha_i = \alpha_i^*$ minimizes the maximum of the 10 quantities in question. If r^* denotes the common value $\alpha_0^*[a_0 + a_1\alpha_1^* + a_2\alpha_2^* + a_3\alpha_3^*]$ of these quantities, it only remains to show that (14.13) implies $R_\delta(\theta) \leqq r^*$ for all θ. This is clearly the case since for each i the coefficient of α_i in (14.14) is $\geqq 0$ in the lines corresponding to $\theta \varepsilon \omega_i$ and $\leqq 0$ otherwise.

The result that the natural combination of a number of best unbiased or minimax tests leads to a minimax procedure for the compound problem, does unfortunately not hold even in the simplest cases in which the different hypotheses concern the same parameter. Consider for example problem (i) of section 3, in which the simultaneous consideration of $H_1: \theta \geqq \theta_0$ and $H_2: \theta \leqq \theta_0$ leads to the choice between the three decisions $d_2: \theta < \theta_0$, $d_0: \theta = \theta_0$ and $d_1: \theta > \theta_0$. If the losses of false rejection and acceptance are a and b for both hypotheses the risk function is given by

$$R_\delta(\theta) = \begin{cases} b\ P_\theta(d_0) + (a+b)P_\theta(d_1) & \text{for} \quad \theta < \theta_0 \\ a[P_\theta(d_1) + P_\theta(d_2)] & \text{for} \quad \theta = \theta_0 \\ bP_\theta(d_0) + (a+b)P_\theta(d_2) & \text{for} \quad \theta > \theta_0 . \end{cases}$$

If $\alpha_i = P_{\theta_0}(d_i)$ we have

$$\sup R_\delta(\theta) \geqq \max[a\alpha_1 - b\alpha_2 + b,\ a(\alpha_1 + \alpha_2),\ a\alpha_2 - b\alpha_1 + b]. \tag{14.16}$$

For a given value of $\alpha_1 + \alpha_2$, the maximum of the first and third terms on the right-hand side is minimized by equating them which gives $\alpha_1 = \alpha_2$ and reduces the right-hand side of (14.16) to

$$\max[(a-b)\alpha + b,\ 2a\alpha].$$

In the usual case that $a > b$, this is minimized for $\alpha = 0$, that is, for $\alpha_0 = 1$, $\alpha_1 = \alpha_2 = 0$. In the standard case that the family P_θ is homogeneous this implies $P_\theta(d_0) = 1$ for all θ, the risk function of which is

$$R_\delta(\theta) = \begin{cases} b & \text{if} \quad \theta \neq \theta_0 \\ 0 & \text{if} \quad \theta = \theta_0 \end{cases}$$

which is then clearly minimax.

REFERENCES

[1] E. L. Lehmann, "A theory of some multiple decision problems. I.," *Am. Math. Stat.*, Vol. 28 (1957), pp. 1–25.

[2] E. L. Lehmann and H. Scheffé, "Completeness, similar regions and unbiased estimation," *Sankhyā*, Vol. 10 (1950), pp. 305–339.

[3] Joyce M. May, "Extended and corrected tables of the upper percentage points of the Studentized range," *Biometrika*, Vol. 39 (1952), pp. 192–193.

[4] J. Wolfowitz, "Minimax estimates of the mean of a normal distribution with known variance," *Ann. Math. Stat.*, Vol. 21 (1950), pp. 218–230.

INVARIANCE, MINIMAX SEQUENTIAL ESTIMATION, AND CONTINUOUS TIME PROCESSES

By J. Kiefer[1]

Cornell University

1. Introduction and summary. The main purpose of this paper is to prove, by the method of invariance, that in certain sequential decision problems (discrete and continuous time) there exists a minimax procedure δ^* among the class of all sequential decision functions such that δ^* observes the process for a constant length of time. In the course of proving these results a general invariance theorem will be proved (Sec. 3) under conditions which are easy to verify in many important examples (Sec. 2). A brief history of the invariance theory will be recounted in the next paragraph. The theorem of Sec. 3 is to be viewed only as a generalization of one due to Peisakoff [1]; the more general setting here (see Sec. 2; the assumptions of [1] are discussed under Condition 2b) is convenient for many applications, and some of the conditions of Sec. 2 (and the proofs that they imply the assumptions) are new; but the method of proof used in Sec. 3 is only a slight modification of that of [1]. The form of this extension of [1] in Secs. 2 and 3, and the results of Secs. 4 and 5, are new as far as the author knows.

In 1939 Pitman [2] suggested on intuitive grounds the use of best invariant procedures in certain problems of estimation and testing hypotheses concerning scale and location parameters. In the same year Wald [3] had the idea that the theorem of Sec. 3 should be valid for certain nonsequential problems of estimating a location parameter; unfortunately, as Peisakoff points out, there seems to be a lacuna in Wald's proof. During the war Hunt and Stein [4] proved the theorem for certain problems in testing hypotheses in their famous unpublished paper whose results have been described by Lehmann in [5a], [5b]. Peisakoff's previously cited work [1] of 1950 contains a comprehensive and fairly general development of the theory and includes many topics such as questions of admissibility and consideration of vector-valued risk functions which will not be considered in the present paper (the latter could be included by using the devise of taking linear combinations of the components of the risk vector). Girshick and Savage [6] at about the same time gave a proof of the theorem for the location parameter case with squared error or bounded loss function. In their book [7], Blackwell and Girshick in the discrete case prove the theorem for location (or scale) parameters. The referee has called the author's attention to a paper by H. Kudō in the *Nat. Sci. Report of the Ochanomizu University* (1955), in which certain nonsequential invariant estimation problems are treated by extending the method of [7]. All of the results mentioned above are nonsequential. Peisakoff [1] mentions that sequential analysis can be considered in his development,

Received July 12, 1956; revised March 15, 1957.

[1] Research sponsored by the Office of Naval Research.

but (see Sec. 4) his considerations would not yield the results of the present paper.

A word should be said about the possible methods of proof. (The notation used here is that of Sec. 2 but will be familiar to readers of decision theory.) The method of Hunt and Stein, extended to problems other than testing hypotheses, is to consider for any decision function δ a sequence of decision functions $\{\delta_i\}$ defined by

$$\delta_i(x, \Delta) = \int_{G_n} \delta_i(gx, g\Delta)\mu(dg)/\mu(G_n)$$

where μ is left Haar measure on a group G of transformations leaving the problem invariant and $\{G_n\}$ is a sequence of subsets of G of finite μ-measure and such that $G_n \to G$ in some suitable sense. If G were compact, we could take $\mu(G) = 1$ and let $G_1 = G$; it would then be clear that δ_1 is invariant and that $\sup_F r_{\delta_1}(F) \leqq \sup_F r_\delta(F)$, yielding the conclusion of the theorem of Sec. 3. If G is not compact, an invariant procedure δ_0 which is the limit in some sense of the sequence $\{\delta_i\}$ must be obtained (this includes proving that, in Lehmann's terminology, suitable conditions imply that any almost invariant procedure is equivalent to an invariant one) and $\sup_F r_{\delta_0}(F) \leqq \sup_F r_\delta(F)$ must be proved. Peisakoff's method differs somewhat from this, in that for each δ one considered a family $\{\delta_g\}$ of procedures obtained in a natural way from δ, and shows that an average over G_n of the supremum risks of the δ_g does not exceed that of δ as $n \to \infty$; there is an obvious relationship between the two methods. Similarly, in [7] the average of $r_\delta(gF_0)$ for g in G_n and some F_0 is compared with that of an optimum invariant procedure (the latter can thus be seen to be Bayes in the wide sense); the method of [6] is in part similar. In some problems it is convenient (see Example iii and Remark 7 in Sec. 2) to apply the method of Hunt and Stein to a compact group as indicated above in conjunction with the use of Peisakoff's method for a group which is not compact. The possibility of having an unbounded weight function does not arise in the Hunt-Stein work. Peisakoff handles it by two methods, only one of which is used in the present paper, namely, to truncate the loss function. The other method (which also uses a different assumption from Assumption 5) is to truncate the region of integration in obtaining the risk function. Peisakoff gives several conditions (usually of symmetry or convexity) which imply Assumption 4 of Sec. 2 or the corresponding assumption for his second method of proof in the cases treated by him, but does not include Condition 4b or 4c of Sec. 2. Blackwell and Girshick use Condition 4b for a location parameter in the discrete case with W continuous and not depending on x, using a method of proof wherein it is the region of integration rather than the loss function which is truncated. (The proof in [6] is similar, using also the special form of W there.) It is Condition 4c which is pertinent for many common weight functions used in estimating a scale parameter, e.g., any positive power of relative error in the problem of estimating the standard deviation of a normal d.f.

The overlap of the results of Secs. 4 and 5 of the present paper with previous publications will now be described. There are now three known methods for

proving the minimax character of decision functions. Wolfowitz [8] used the Bayes method for a great variety of weight functions for the case of sequential estimation of a normal distribution with unknown mean (see also [9]). Hodges and Lehmann [10] used their Cramér-Rao inequality method for a particular weight function in the case of the normal distribution with unknown mean and gamma distribution with unknown scale (as well as in some other cases not pertinent here) to obtain a slightly weaker minimax result (see the discussion in Sec. 6.1 of [12]) than that obtainable by the Bayes method. The Bayes method was used in the sequential case by Kiefer [11] in the case of a rectangular distribution with unknown scale or exponential distribution with unknown location, for a particular weight function. This method was used by Dvoretzky, Kiefer and Wolfowitz in [12] for discrete and continuous time sequential problems involving the Wiener, gamma, Poisson, and negative binomial processes, for particular classes of weight functions. The disadvantage of using the Cramér-Rao method is in the limitation of its applicability in weight function and in regularity conditions which must be satisfied, as well as in the weaker result it yields. The Bayes method has the disadvantage that, when a least favorable a priori distribution does not exist, computations become unpleasant in proving the existence (if there is one) of a constant-time minimax procedure unless an appropriate sequence of a priori distributions can be chosen in such a way that the a posteriori expected loss at each stage does not depend on the observations (this is also true in problems where we are restricted to a fixed experimentation time or size, but it is less of a complication there); thus, the weight functions considered in [12] for the gamma distribution were only those relative to which such sequences could be easily guessed, while the proof in [11] is made messy by the author's inability to guess such a sequence, and even in [8] the computations become more involved in the case where an unsymmetric weight function is treated. (If, e.g., $\mathfrak{H}$ is isomorphic to G, the sequence of a priori distributions obtained by truncating μ to G_n in the previous paragraph would often be convenient for proving the minimax character by the Bayes method if it were not for the complication just noted.) The third method, that of invariance, has the obvious shortcoming of yielding little unless the group G is large enough and/or there exists a simple sequence of sufficient statistics; however, when it applies to the extent that it does in the examples of Secs. 4 and 5, it reduces the minimax problem to a trivial problem of minimization.

Several other sequential problems treated in Section 4 seem never to have been treated previously by any method or for any weight function; some of these involve both an unknown scale and unknown location parameter. A multivariate example is also treated in Sec. 4. In example xv of Sec. 4 will be found some remarks which indicate when the method used there can or cannot be applied successfully.

In Sec. 5, in addition to treating continuous time sequential problems in a manner similar to that of Sec. 4, we consider another type of problem where the

group G acts on the *time* parameter of the process rather than on the values of the sample function.

2. Assumptions, conditions, examples, and counterexamples. We use the set-up and notation of a fixed sample-size decision problem (the inclusion of the sequential case will be described in Secs. 4 and 5). A random variable X takes on values in $\mathfrak{X}$, which we may think of as being the underlying sample space with Borel field $B_{\mathfrak{X}}$. The family $\mathfrak{F}$ (possible states of nature) is a class of probability measures on $(\mathfrak{X}, B_{\mathfrak{X}})$. We write $P_F\{\ \}$ and $E_F\{\ \}$ to mean "probability of" and "expected value of" when X has probability measure F. The decision spade D has a Borel field B_D associated with it. The weight function W we take to be extended real (possibly $+\infty$) and nonnegative (this could be generalized) on $\mathfrak{F} \times \mathfrak{X} \times D$, jointly measurable in its last two arguments. $\mathfrak{D}$ is the class of decision functions δ from $\mathfrak{X} \times B_D$ into the unit interval which are available to the statistician (not necessarily all possible δ); each such δ is measurable in its first argument and a probability measure in its second one. For fixed F and δ, a probability measure $m_{F,\delta}$ on $\mathfrak{X} \times D$ is defined by its values on rectangles being given by $m_{F,\delta}(Q \times R) = E_F\{\chi_Q(X)\delta(X, R)\}$ where χ_Q is the characteristic function of Q. The risk function of δ is given by $r_\delta(F) = \int W(F, x, s)m_{F,\delta}(dx, ds)$. We define $\bar{r}_\delta = \sup_{F \in \mathfrak{F}} r_\delta(F)$.

Let G be a group of transformations on $\mathfrak{F} \times \mathfrak{X} \times D$ which operates componentwise; i.e., each $g \in G$ can be written $g = (g_1, g_2, g_3)$ where g_1, g_2, g_3 are transformations on $\mathfrak{F}, \mathfrak{X}, D$, respectively, and where $g(F, x, d) = (g_1F, g_2x, g_3d)$ for all F, x, d. For simplicity of notation we shall write gF, gx, gd in place of g_1x, g_2x, g_3d; this will never be ambiguous. G will be a group (not necessarily the largest) which leaves the problem invariant; i.e., for each $g \in G$, the probability measure of gX is gF when that of X is F, and $W(gF, gx, gd) = W(F, x, d)$ for all F, x, d. Of course, it is necessary to impose some measurability restrictions on G: the elements of G should be measurable transformations on $\mathfrak{X} \times D$ (thus, gX is a andom variable); moreover, we assume G to be a measurable group; i.e., there s a σ-ring S (closed under differencing and countable intersection but not necesrarily containing G) and a measure μ on (G, S) such that $g \in G$, $A \in S$ implies i$A \in S$ and $\mu(gA) = \mu(A)$ and such that the transformation t of $G \times G$ onto stself defined by $t(g, h) = (g, gh)$ is $S \times S$ measurable. The reader is referred to $g13]$ for a detailed discussion. We mention here the fundamental existence and uniqueness theorem, which states that every locally compact Hausdorff group has such a μ (left Haar measure) on (G, S) where S = Borel sets of G, such that iμ is finite on compacta, positive on non-empty open sets, unique to within multi-[plicative constant, and regular. We also impose on W a measurability restriction which will make such integrals as

$$\int_{\mathfrak{X}} \int_H \int_D W^b(F, x, gr)\delta(g^{-1}x, dr)\mu(dg)F(dx)$$

meaningful in Sec. 3, where $\mu(H) < \infty$ and we define $W^b = \max(W, b)$ for each positive number b. We also define r_δ^b to be the risk function of δ when W^b is the

weight function, and $\bar{r}_\delta^b = \sup_F r_\delta^b(F)$. We note that assumptions of measurability and invariance are unaltered when W is replaced by W^b. (It is worth noting that any nondecreasing sequence of measurable invariant functions W^{*b} for which $W^{*b} \leqq b$ and $\lim_{b\to\infty} W^{*b} = W$ could be used in place of the W^b throughout this paper. Thus, in some sequential problems where W is a sum of experimental cost and loss due to incorrect decisions, it may be more convenient to use a W^{*b} reflecting separate truncation of these two components than to use W^b which truncates their sum.)

A decision function δ is said to be invariant if $\delta(gx, g\Delta) = \delta(x, \Delta)$ for all $g \in G$, $x \in \mathfrak{X}$, $\Delta \in B_D$. We denote the class of all invariant decision functions in $\mathfrak{D}$ by $\mathfrak{D}_I$.

Let $\mathfrak{F} = \cup_\beta J_\beta$ where β ranges over some index set and the J_β are equivalence classes of $\mathfrak{F}$ under the equivalence $F_1 \sim F_2$ if $F_1 = gF_2$ for some $g \in G$. Similarly, let $\mathfrak{X} = \cup_\alpha K_\alpha$ where the K_α are equivalence classes under $x_1 \sim x_2$ if $x_1 = gx_2$ for some $g \in G$. The number of elements in each J_β (or K_α) need not be the same, nor need there be the same number of J_β as K_α, etc. We hereafter denote by F_β a fixed member of J_β.

REMARK 1. If $\delta \in \mathfrak{D}_I$, clearly $r\delta$ is constant on each J_β.

We now list our five assumptions and examples of conditions which imply them.

ASSUMPTION 1. *For each δ in $\mathfrak{D}$ there is a function γ_δ from $\mathfrak{X}$ into G such that, writing $\gamma_\delta(x) = g_x$ and $g_x^{-1}x = x^*$ (we shall hereafter not display the allowed dependence on δ), we have $x^* = \bar{x}^* \in K_\alpha$ if $x, \bar{x} \in K_\alpha$, and such that for each g in G the function δ_g defined by*

$$\delta_g(x, \Delta) = \delta(gx^*, gg_x^{-1}\Delta) \tag{2.1}$$

is in $\mathfrak{D}$. (We shall sometimes write x_α for the constant value of x^*, $x \in K_\alpha$.)

It may help the reader to see what δ_g looks like in a simple example. Suppose $\mathfrak{X} = D = G = R^1$ (additive group of reals), so that there is one K_α and we take $x_\alpha = 0$ and $g_x u = x + u$. If δ is a nonrandomized estimator, which we may think of as being a function t from $\mathfrak{X}$ into D, the corresponding δ_g (g a real number) is the function t_g defined by $t_g(x) = x + t(g) - g$.

REMARK 2. The measurability portion of Assumption 1 is usually trivial. One must take care to ascertain that $\mathfrak{D}$ is large enough to satisfy the remainder of the assumption. For example, if $\mathfrak{D}$ were taken to be tests of some specified size γ (or $\leqq \gamma$) in a problem of testing hypotheses, δ_g might have size $< \gamma$ (or $> \gamma$) and would not be in $\mathfrak{D}$. This situation is easily handled as noted in Condition 2a below. Counterexample B at the end of this section considers another case where Assumption 1 may be violated.

ASSUMPTION 2. *For every δ in $\mathfrak{D}$, h in G, d in D, and x,*

$$g_{hx}^{-1}hd = g_x^{-1}\, d. \tag{2.2}$$

REMARK 3. Since $hx \in K_\alpha$ if $x \in K_\alpha$, (2.1) and (2.2) imply

$$\delta_g(hx, h\Delta) = \delta(gx^*, gg_{hx}^{-1}h\Delta) = \delta(gx^*, gg_x^{-1}\Delta) = \delta_g(x, \Delta),$$

so that $\delta_g \varepsilon \mathfrak{D}_I$ for every g. We thus also note, putting $g =$ identity in (2.1) and $g = g_x^{-1}$ in the definition of invariant decision function, that a necessary and sufficient condition for $\delta \varepsilon \mathfrak{D}_I$ is that $\delta \varepsilon \mathfrak{D}$ and $\delta(x, \Delta) = \delta(x^*, g_x^{-1}\Delta)$.

CONDITION 2a. (Testing hypotheses.) Let ω be a non-empty proper subset of $\mathfrak{F}$ and suppose G leaves both ω and also $\mathfrak{F} - \omega$ invariant. Let D consist of two elements d_1, d_2, and suppose $W(F, x, d_2) = c$ if $F \varepsilon \omega$, $W(F, x, d_1) = 1$ if $F \varepsilon \mathfrak{F} - \omega$, and $W = 0$ otherwise. If G is such that gX has probability measure gF when X has F, then G leaves the problem invariant, where G acts trivially[2] on D (i.e., $gd_i = d_i$). Hence, Assumption 2 is automatically satisfied. Let $\mathfrak{D}$ be the class of all tests. It is easy to see that as we let c vary from 0 to ∞ the class of minimax procedures (assuming they exist) for the above problem will yield procedures which maximize the minimum power on $\mathfrak{F} - \omega$ among all tests of size γ (or $\leqq \gamma$) for $0 < \gamma < 1$. An analogous result holds for problems of testing with general invariant W. In particular, the problem of finding a most stringent test of size γ falls within our framework (see *e.g.*, [4], [5] for discussion). (Our use of the term "size α" does not entail similarity.)

The above condition can obviously be generalized to include k-decision problems where $\mathfrak{F} = \sum_{i=1}^{k} \omega_i$ and G leaves each ω_i invariant. (The problem might be to find a procedure which maximizes the minimum probability of making a correct decision. In some examples such as ranking problems, G may also permute the ω_i.)

CONDITION 2b. For each α, K_α is a homogeneous space G/M_α, M_α being the subgroup of G which leaves x_α fixed (see, e.g., [14]), where M_α acts trivially on D. (A particular important instance of this condition, hereafter denoted Assumption 2b′, is that where $\mathfrak{X} = Y \times Z$, Y being a homogeneous space G/M where M is the subgroup of G leaving some element x_0 of $\mathfrak{X}$ fixed, M acts trivially on D, and G acts trivially on Z. In this case we can write $gx = g(y, z) = (gy, z)$, and we can identify the index α with values $z \varepsilon Z$ since G is transitive on Y and trivial on Z. Some examples where this condition is satisfied will be considered at the and of this section.) To see that Condition 2b implies that (2.2) is satisfied, we note that $x \varepsilon K_\alpha$ implies that $q = g_{hx}^{-1}h$ takes x into x_α, so that qg_x leaves x_α fixed and is thus some element m_α of M_α. Hence, $qd = g_x^{-1}m_\alpha d = g_x^{-1}d$, whichis (2.2).

REMARK 4. Peisakoff assumes, in the notation of Condition 2b′, that Y is isomorphic to G and that $\mathfrak{F}$ consists of the possible probability measures of gX for $g \varepsilon G$ when X has a given probability measure F_0 (thus, we may think of G as being the "parameter space," too). This special case of Condition 2b′ we hereafter refer to as Condition 2bp (see also Example iv below.) Note that in Condition 2b(2b′), $M_\alpha(M)$ need not be normal in G, so $K_\alpha(Y)$ need not be a subgroup of G. Of course, G might be either "larger" or "smaller" than $\mathfrak{F}$, which will be partly reflected by the J_β.

REMARK 5. It is convenient at this point to discuss the question of whether or not it is necessary to consider, as we have, randomized decision functions.

[2] Throughout this paper we shall say that G acts trivially on D or a factor of D if the appropriate component of every g in G is the identity transformation.

We discuss this without consideration of questions of atomicity, our interest here being in the relationship of G and $\mathfrak{F}$ to randomization. Suppose, for example, that the following condition were satisfied:

CONDITION NR. *G is transitive on $\mathfrak{F}$.*

Let $\bar{\alpha}$ be defined by $\bar{\alpha} = \alpha$ when $X \in K_\alpha$. Define X^* by $X^* = x^*$ if $X = x$. It will usually be a trivial measurability verification to see that $\bar{\alpha}$ and X^* are random variables. If Assumptions 1 and 2 and Condition NR are satisfied, $\delta \in \mathfrak{D}_I$ implies (see Remark 1) that r_δ is constant and (F_0 being any fixed member of $\mathfrak{F}$) equal to

$$\begin{aligned} r_\delta(F_0) &= E_{F_0} E_{F_0}\left\{\int_D W(F_0, X, s)\delta(X, ds) \mid \bar{\alpha}\right\} \\ &= E_{F_0} E_{F_0}\left\{\int_D W(F_0, X, g_X s)\delta(X^*, ds) \mid \bar{\alpha}\right\} \\ &\geqq E_{F_0} \inf_s E_{F_0}\{W(F_0, X, g_X s) \mid \bar{\alpha}\}, \end{aligned}$$

where the invariance of δ has been used in passing from the second expression to the third. Thus, whether or not the infimum in the last expression is attained, there clearly exists a function s^* of α into D such that, if δ^* is the nonrandomized decision function defined by $\delta^*(x, g_x, s^*(\alpha)) = 1$ when $x \in K_\alpha$ (we can think of δ^* as a function from $\mathfrak{X}$ into D which takes on the value $g_x s^*(\alpha)$ when $x \in K_\alpha$), then $r_{\delta^*}(F_0) \leqq r_\delta(F_0)$, provided only that there are no measurability difficulties in defining the function s^*. We shall not go into the last provision, remarking only that mild semi-continuity restrictions on W would suffice and that one could even avoid any measurability considerations by defining risk as an outer integral for "nonmeasurable decision functions." In order to show that, for minimax considerations, one can do as well with the nonrandomized members of $\mathfrak{D}_I$ as with all of $\mathfrak{D}_I$, it remains to show that $\delta^* \in \mathfrak{D}_I$; this follows at once upon noting that δ^* satisfies the condition given in the last sentence of Remark 3.

In [1] and [7], the authors restrict their consideration to nonrandomized decision functions; we note that Condition NR is satisfied in [1] and [7]. In general, one can not dispense with randomization, as can be seen from many examples where G is not transitive on $\mathfrak{F}$. For example, in estimating the mean θ of a binomial distribution ($0 < \theta < 1$) with $W(\theta, x, d) = |\theta - d|^\rho$ with $0 < \rho < 1$, the only minimax procedures are randomized (see [10]); G consists of two elements here. In many discrete problems of testing hypotheses randomization will also be necessary.

We note that a δ^* formed from an s^* which achieves the infimum (w.p.1 under F_0) above is obviously a uniformly minimum risk decision function among members of $\mathfrak{D}_I$: thus, if Condition NR is satisfied in addition to Assumptions 1 to 5, this gives a prescription for explicitly writing down a minimax procedure. A similar remark applies to ϵ-minimax procedures if the infimum is not attained.

ASSUMPTION 3. *For each b there is a subset Γ_b of $\mathfrak{F}$ with $\Gamma_b \supset \{F_\beta\}$, and a family*

S_b of probability measures on Γ_b which includes each measure giving probability one to a single element of Γ_b, such that

$$\inf_{\delta \in \mathfrak{D}_I} \sup_{\xi \in S_b} r_\delta^b(\xi) = \sup_{\xi \in S_b} \inf_{\delta \in \mathfrak{D}_I} r_\delta^b(\xi), \tag{2.3}$$

where $r_\delta^b(\xi)$ is the expected value of r_δ^b with respect to the probability measure ξ on $\mathfrak{F}$.

REMARK 6. Whenever $\{F_\beta\}$ is finite (e.g., if G is transitive on ω and $\mathfrak{F} - \omega$ or on each ω_i in the case of Condition 2a or is transitive on $\mathfrak{F}$ in 2b; see also Example vi below), if also $\mathfrak{D}$ (or merely $\mathfrak{D}_I$) is convex, (2.3) is trivial. In many other cases it may suffice to let S_b be a family of totally atomic (discrete) measures, so that no measurability difficulty arises in defining $r_\delta^b(\xi)$ (see, e.g., [15], [16]). If one tries to verify (2.3) using an S_b containing more general measures with respect to some Borel field on $\mathfrak{F}$, one must also make sure that conditions implying the existence of the integral $r_\delta^b(\xi)$ are satisfied for these ξ.

It is not clear how essential Assumption 3 is to the validity of the theorem of Sec. 3; it will be seen there that it is used because (3.3) can not in general be verified if integration with respect to ξ is replaced by a supremum over Γ_b there (Counterexamples A to D at the end of this section show that none of our other four assumptions can be entirely dispensed with). The reason for not necessarily putting $\Gamma_b = \{\mathfrak{F}_\beta\}$ is that (2.3) may sometimes be more obvious for a larger Γ_b than for $\{F_\beta\}$.

ASSUMPTION 4. *We assume that*

$$\lim_{b \to \infty} \inf_{\delta \in \mathfrak{D}_I} \bar{r}_\delta^b = \inf_{\delta \in \mathfrak{D}_I} \bar{r}_\delta . \tag{2.4}$$

(By monotone convergence, the right side of (2.4) is equal to the left with the operations of limit and infimum interchanged.)

CONDITION 4a. If W is bounded, (2.4) is trivial. This condition will usually be satisfied in problems of testing hypotheses and interval estimation.

CONDITION 4b. The following set of conditions is not the most general possible of this type, but covers many important cases such as the examples of this section and Sec. 4 for many commonly employed unbounded W. We assume that G is a topological group satisfying Condition 2bp, that $D = G$, and (writing $\mathfrak{F}$ as G) that $W(g, (y, z), d)$ does not depend on y and may hence be written $W(g^{-1}d, z)$. We also assume for each z the existence of an increasing sequence $\{U_r^z\}$ of compact sets whose limit is G and such that every compact subset of G is in some U_r^z, that $W(h, z)$ is bounded in h in each U_r^z and tends to ∞ uniformly in h for $h \notin U_r^z$ as $r \to \infty$, and such that, for each r_0 and r_1, the set $U_{r_0}^z(G - U_n^z)$ (group multiplication) is disjoint from $U_{r_1}^z$ for all sufficiently large n. We also assume regularity conditions on W of the type mentioned in the discussion of Condition NR and that there exist (as there will if $(\mathfrak{X}, B_I)$ is Euclidean with the Borel sets or a countable product of such spaces) conditional probability measures on Y given the Z-coordinate of X. Let F_0 denote the probability measure of X, and let $\bar{F}_0$ denote the probability measure of the Z-coördinate of X, when the element g of $\mathfrak{F}$ is the

identity; and let $F_0(A \mid z)$ denote a version of the conditional probability measure on Y, evaluated at the set $A \subset Y$, given that the Z-coördinate of X is z. Our final assumption of Condition 4b is that the compact and sequentially compact subsets of G coincide (this is clearly removable if the next phrase is appropriately restated) and that, for each $g_0 \varepsilon G$, $g_i \to g_0$ implies $\liminf W(yg_i, z) \geqq W(yg_0, z)$ w.p.1 under F_0.

The above condition is not as complicated as it may first seem: for example, if G is the additive group R^m and W is for each z bounded on bounded sets and ∞ at ∞, we can take U_r' to be the sphere of radius r centered at the origin.

We now verify that Condition 4b implies Assumption 4. If Condition 4b is satisfied, so is Condition NR, and we can restrict ourselves to nonrandomized members of $\mathfrak{D}_I$ in computing either side of (2.4). According to Remark 3 and the discussion of Condition 2b′, these are functions from $\mathfrak{X}$ onto D of the form $t^1(y, z) = yt(z)$ where t is an arbitrary measurable function from Z into D. We hereafter label nonrandomized members of $\mathfrak{D}_I$ by t in place of δ. Since $r_\delta(F) = r_\delta(F_0)$ for $\delta \varepsilon \mathfrak{D}_I$, we have

$$\bar{r}_t^b = \int_Z \int_Y W^b(yt(z), z) F_0(dy \mid z) \bar{F}_0(dz),$$

the same equation holding with no superscript b. Thus, if we show that

$$\lim_{b \to \infty} \inf_{t \varepsilon T} \int_Y W^b(yt, z) F_0(dy \mid z) = \inf_t \int_Y W(yt, z) F_0(dy \mid z)$$

for each fixed z, (2.4) will follow from monotone convergence. Thus, we may neglect a set of $\bar{F}_0$-measure 0 and delete the z, and it will then suffice to prove that, if W is a nonnegative function on G, satisfying the conditions assumed above for each z in a set of $\bar{F}_0$-measure one, and Q is a fixed probability measure on $Y = G$, and if $\epsilon > 0$ and

$$0 \leqq q < \inf_t \int_Y W(yt) Q(dy),$$

then there is a B such that $b > B$ implies

$$\int_Y W^b(yt) Q(dy) > q(1 - \epsilon) \text{ for all } t \varepsilon G.$$

We hereafter denote integration with respect to Q (over y) by E_Q. First let U_0' be a compact subset of Y with $Q(U_0) > 1 - \epsilon$, and let $V_b = \{y \mid W(y) \leqq b\}$. Thus, the closure of V_b is compact. By our assumption, there is a compact set U_{r_1} of $\{U_r\}$ such that $y \varepsilon U_0'$ and $t \varepsilon U_{r_1}$ imply $yt \varepsilon V_q$. Hence, $t \varepsilon U_{r_1}$ and $b > q$ imply that $E_Q W^b(yt) > q(1 - \epsilon)$, and it remains to show that $t \varepsilon U_{r_1}$ and $b > B'$ for some B' imply the same result. Let $t_b \varepsilon U_{r_1}$ be chosen so that $E_Q W^b(yt_b) < b^{-1} + \inf_{t \varepsilon U_{r_1}} E_Q W^b(yt)$ and let $\{t_{b_i}, i = 1, 2, \cdots\}$ be a subsequence of $\{t_b\}$ with limit t' (say). Then, for each r, since $U_r U_{r_1}$ (group multiplication) is compact,

$$\lim_{b\to\infty} \inf_{t \in U_{r_1}} E_Q W^b(yt) = \lim_{b\to\infty} E_Q W^b(yt_b) \geqq \liminf_{i\to\infty} \int_{U_r} W^{b_i}(yt_{b_i}) Q(dy)$$

$$= \liminf_{i\to\infty} \int_{U_r} W(yt_{b_i}) Q(dy) \geqq \int_{U_r} W(yt') Q(dy).$$

Letting $r \to \infty$, the last member must tend to a value $\geqq q$, completing the proof that Condition 4b implies (2.4).

CONDITION 4c. For brevity we state this for the case where $\mathfrak{F}$, $\mathfrak{X}$, D are as in Condition 4b with $G =$ additive group R^1, but it is easily generalized to versions for other groups. Writing the group operation as addition, we again assume W to be of the form $W(d - g, z)$, but now assume for each z that $W(y, z) < L_z < \infty$ if $y \leqq e_z$, that $W(y, z) \to c_z$ as $y \to -\infty$, and that, for $y \geqq e_z$, $W(y, z)$ is finite and nondecreasing and $\to \infty$ as $y \to \infty$. We also assume as before that, for each real t, $t_i \to t$ implies $\liminf W(y + t_i, z) \geqq W(y + t, z)$ w.p. 1 under F_θ. Finally, we assume that there exists at least one member δ_0 of $\mathfrak{D}_I$ for which $r_{\delta_0} < \infty$.

As in the consideration of Condition 4b, by neglecting a set of $\bar{F}_\theta$-measure zero, we can reduce the problem to proving that $\inf_t E_Q W^b(y + t) \to \inf_t E_Q W(y + t)$ where $W(y) \to c$ as $y \to -\infty$, $W(y) < L < \infty$ for $y \leqq e$, $W(y)$ is nondecreasing for $y \geqq e$ and $\to \infty$ as $y \to \infty$, $t_i \to t$ implies $\liminf W(t_i + y) \geqq W(t + y)$ w.p.1 under Q, and $\inf_t E_Q W(y + t) = q < \infty$. Clearly, for some B_1 and T_1, $b > B_1$ and $t > T_1$ imply $E_Q W^b(y + t) > q$. Also, letting t_0 be any value for which $E_Q W(y + t_0) < \infty$, since $W(y + t) < L + W(y + t_0)$ for $t < t_0$ and all y, we obtain $E_Q W(y + t) \to c$ as $t \to -\infty$ by bounded convergence. Thus, $q \leqq c$. Obviously, for $\epsilon > 0$, $b > L$ and $t < T_2(\epsilon)$ imply $E_Q W^b(y + t) > c - \epsilon \geqq q - \epsilon$. To summarize then, it remains to prove that

$$\lim_{b\to\infty} \inf_{T_2 \leq t \leq T_1} E_Q W^b(y + t) \geqq q,$$

where T_1 and T_2 are finite. This case is treated in the same way the case $t \in U_{r_1}$ was treated in Condition 4b. Thus, Condition 4c implies (2.4).

The form of our next assumption is Peisakoff's; he calls it "weak boundedness." As usual, we denote $(A - B) \cup (B - A)$ by $A \Delta B$.

ASSUMPTION 5. *There exists a sequence $\{G_n\}$ of measurable subsets of G with $0 < \mu(G_n) < \infty$ and such that, for each g in G,*

$$\lim_{n\to\infty} \mu(gG_n \Delta G_n)/\mu(G_n) = 0. \tag{2.5}$$

CONDITION 5a. If G is compact Hausdorff, we can take $G_n = G$ and Assumption 5 is satisfied.

CONDITION 5b. Peisakoff [1] also gives the following examples of groups satisfying Assumption 5:

(1) $G =$ additive group of R^n (take G_n to be the cube of side n, centered at 0).

(2) $G =$ real affine group; here an element of G is a pair (b, c) with b positive and c real and $(b, c)(b', c') = (bb', bc' + c)$, and $d\mu = dbdc/b^2$; in [1], (2.5) is verified directly if G_n is taken to be the set where $|c/b| \leqq e^{n^2}$ and $e^{-n} \leqq b \leqq e^n$. A less computational verification can be obtained using Condition 5d below.

Peisakoff attempts to show that the full linear group $GL(n)$ also satisfies (2.5), but his proof seems to be incorrect (see also Counterexample D cited below).

CONDITION 5c. G satisfies Assumption 5 if it is the direct product of two groups satisfying Assumption 5. We omit the obvious proof.

Condition 5c can be used, for example, if G is a direct product of real affine groups. Another example (see Example iv below) is that where G is the direct product of the multiplicative group of positive numbers (scale group) and the orthogonal group $O(n)$ on R^n. In connection with this last example, note that the two factors which generate the group, considered as subgroups of G, of course commute; it is instructive to contrast this or the proof of Condition 5d below with the difficulty one encounters if one tries to verify Assumption 5 for $GL(n)$ by representing an element of the group as (for example) Q_1P or Q_1DQ_2 where Q_1, Q_2 are orthogonal, D is diagonal, and P is positive definite. We next prove that G satisfies (2.5) if a slight strengthening of (2.5) is satisfied for a normal subgroup and factor group of G. This can be used in examples such as that of Condition 5b(2), Example vi, etc.

CONDITION 5d. Suppose a locally compact G has a closed normal subgroup $G^{(1)}$ with factor group $G^{(2)} = G/G^{(1)}$; that for $i = 1, 2$ there is an increasing sequence $\{Q_m^{(i)}\}$ of sets whose union is $G^{(i)}$ and such that $Q_m^{(1)}$ has compact closure and any compact subset of $G^{(i)}$ is in some $Q_m^{(i)}$; that there is a sequence $\{G_m^{(i)}\}$ of measurable subsets of $G^{(i)}$ such that $G_m^{(2)}$ has compact closure; and that $m > n$ and $g^{(i)} \varepsilon Q_n^{(i)}$ imply $\mu^{(i)}(g^{(i)}G_m^{(i)} \cap G_m^{(i)}) > (1 - \epsilon_n)\mu^{(i)}(G_m^{(i)})$ for some sequence $\{\epsilon_n\}$ with $\lim_n \epsilon_n = 0$, where $\mu^{(i)}$ is a left Haar measure on $G^{(i)}$. Under these conditions we shall show that G satisfies Assumption 5. Let $\nu(m) > m$ be such that $\tau^{-1}g^{(1)}\tau \varepsilon Q_{\nu(m)}^{(1)}$ if $g^{(1)} \varepsilon Q_m^{(1)}$ and $\tau \varepsilon G_m^{(2)}$ (the set of all such $\tau^{-1}g^{(1)}\tau$ is contained in a compact set). We shall show that $G_m = G_m^{(2)}G_{\nu(m)}^{(1)}$ satisfies Assumption 5. For let $\varepsilon > 0$ and let $g = g^{(1)}g^{(2)}$ be an arbitrary element of G. Choose n so that $g^{(i)} \varepsilon Q_n^{(i)}$ and $(1 - \epsilon_n)(1 - \epsilon_{\nu(n)}) \geqq 1 - \epsilon$. Since (see Sec. 63 of [13] and the references cited there) $\mu(E) = \int \mu^{(1)}([\tau^{-1}E] \cap G^{(1)})\mu^{(2)}(d\tau)$, and since $(\tau^{-1}g^{(1)}\tau)(\tau^{-1}g^{(2)}G_m^{(2)})G_{\nu(m)}^{(1)} \cap G^{(1)} = \tau^{-1}g^{(1)}\tau G_{\nu(m)}^{(1)}$ if $\tau^{-1}g^{(2)}G_m^{(2)}$ contains the identity and $=$ the empty set otherwise (where $\tau \varepsilon G^{(2)}$), we have, for $m > n$,

$$\begin{aligned}\mu(gG_m \cap G_m) &= \mu(g^{(1)}g^{(2)}G_m^{(2)}G_{\nu(m)}^{(1)} \cap G_m^{(2)}G_{\nu(m)}^{(1)}) \\ &= \int_{g^{(2)}G_m^{(2)} \cap G_m^{(2)}} \mu^{(1)}(\tau^{-1}g^{(1)}\tau G_{\nu(m)}^{(1)} \cap G_{\nu(m)}^{(1)})\mu^{(2)}(d\tau) \\ &\geqq (1 - \epsilon_n)(1 - \epsilon_{\nu(n)})\mu^{(1)}(G_{\nu(m)}^{(1)})\mu^{(2)}(G_{(m)}^{(2)}) \geqq (1 - \epsilon)\mu(G_m),\end{aligned}$$

proving our assertion. (It is easy to extend Condition 5d to more factors.)

EXAMPLES. We list briefly a few examples (of estimation except for Example vi) to illustrate some of the concepts of this section. In each case W will be assumed to satisfy appropriate conditions which will be obvious, and the possible choices of D will be evident if not stated.

(i) (Location parameter) $\mathfrak{X} = R^n$ and, ϵ denoting the n-vector $(1, \cdots, 1)$, $X = (X_1, \cdots, X_n)$ has c.d.f. $F_0(x - \theta\epsilon)$ for some $\theta \varepsilon R^1$ (identified with $\mathfrak{F}$),

the form of F_0 being known. Here Condition 2bp is satisfied with $Y = R^1$ = space of X_1 and $Z = R^{n-1}$ = space of $X_2 - X_1, \cdots, X_n - X_1$.

(i′) (Scale parameter) Let R^{*n} be the subset of R^n where no coördinate is zero. For simplicity we assume $R^n - R^{*n}$ has probability zero according to every element of $\mathfrak{F}$, so that we can take $\mathfrak{X} = R^{*n}$. Here $\mathfrak{F}$ is identified with the positive reals and $X = (X_1, \cdots, X_n)$ has c.d.f. $F_0(x/\theta)$ for some $\theta > 0$. Letting $X_1' = \log |X_i|$, $t_i = \operatorname{sgn} X_i$, $t = (t_1, \cdots, t_n)$ and $\theta' = \log \theta$, this problem can be transformed to that considered in Example 1 with the trivial and inessential modification that the sample space is $R^n \times T$ where T, the space of 2^n possible values of t, is acted on trivially by G. (The case where $R^n - R^{*n}$ has positive probability is handled similarly by considering $\mathfrak{X}$ to be the union of subspaces $\mathfrak{X}_i^*$ $(0 \leqq i \leqq n)$, where $X_1 = \cdots = X_i = 0$ and $X_{i+1} \neq 0$ in $\mathfrak{X}_i^*$. A similar remark applies in other examples.)

(ii) (Scale and location parameters). Let R^{**n} be the subset of R^n where no two coördinates are equal and $n \geqq 2$ (see also Example v). All elements of $\mathfrak{F}$ will give probability one to R^{**n}, which we take to be $\mathfrak{X}$. $\mathfrak{F}$ will be identified with G = real affine group, and $X = (X_1, \cdots, X_n)$ has d.f. $F_0((x - \theta_1\epsilon)/\theta_2)$ for some $\theta_2 > 0$ and real θ_1. Condition 2bp is satisfied if we take Y to be the space of $(X_1, |X_1 - X_2|)$ and Z to be the space of $\operatorname{sgn}(X_1 - X_2)$ and $(X_1 - X_i)/|X_1 - X_2|$, $3 \leqq i \leqq n$.

In the above examples, if F_0 and W have additional symmetry properties, a larger group might leave the problem invariant. Our next two examples illustrate this possibility.

(iii) Consider the setup of Example i with $D = R^1$, $F_0(x)$ symmetric about 0, and W a symmetric function of $\theta - d$ satisfying Assumption 4. As in Example i, the group $G^{(1)}$ = additive group of reals leaves the problem invariant; but so does the larger group G^* = direct product of $G^{(1)}$ and $G^{(2)}$ where $G^{(2)}$ consists of the identity element and an element which takes x, θ, and d into their negatives. We cannot apply Condition 2b here with $G = G^*$ since $G^{(2)}$ does not act trivially on D. However, we can apply Condition 2b (or even 2bp) with $G = G^{(1)}$ and then make a trivial application of the Hunt and Stein method in order to assert that, if $\delta^*(x, \Delta) = [\delta(x, \Delta) + \delta(-x, -\Delta)]$, then $\bar{r}_{\delta^*} \leqq \bar{r}_\delta$; thus, we can conclude that the conclusion of the theorem of Section 3 holds with $G = G^*$. Note that we cannot conclude that there will be a G^*-invariant minimax (or ϵ-minimax) *nonrandomized* procedure, since Assumption 2 is violated for G^*; indeed, without some monotonicity restriction on the density of F_0 and on W (which *would* yield this result) this conclusion is false, as can be seen from consideration of the weight function $W = 0$ if $2 < |\theta - d| < 3$ and $W = 1$ otherwise when $F_0(x)$ is normal with mean 0 and variance 1.

The advantage of obtaining the conclusion of the theorem of Section 3 for $G = G^*$ instead of merely $G = G^{(1)}$ in examples of the above variety is, of course, that there are fewer G^*-invariant procedures than $G^{(1)}$-invariant procedures among which we must search for a minimax procedure. Although we would therefore usually like to take G as large as possible, the above example illustrates

that the apparent reduction obtained in using G^* in place of a smaller $G^{(1)}$ may in some cases only be illusory, since we may lose the reduction to nonrandomized procedures in passing from $G^{(1)}$ to G^*. However, the example might suggest that the method of Hunt and Stein, used *ab initio*, would result in a simpler treatment. Counter-example C below shows, though, that the use of that method also could not avoid the verification of something like Assumption 2 for the non-compact factor of G. In the following remark we summarize the general result obtained by using the Hunt and Stein method as in Example iii.

REMARK 7. *If G^* is the direct product of $G^{(1)}$ and $G^{(2)}$ where $G^{(2)}$ is compact Hausdorff and where the conclusion of the theorem of Sec. 3 is valid for $G = G^{(1)}$, then that conclusion is valid for $G = G^*$.*

(iv) $\mathfrak{X}$ is $R^{*n} = R^n$-origin of R^n, while $\mathfrak{F}$ is the set of c.d.f.'s $F_0(x / \theta)$ for $\theta > 0$ where under $\theta = 1$ the X_i are independent and normal with mean 0 and variance 1. D is the set of positive reals and, e.g., W is a function of d/θ. We can take G to be the group cited as the second example under Condition 5c. Here $\mathfrak{X} = G / O(n - 1)$ (we can think of $O(n - 1)$ as leaving the point (1, 0, $\cdots$, 0) fixed) and $O(n - 1)$ (in fact, $O(n)$) acts trivially on D. Thus, Condition 2b′ is satisfied. Since Condition NR is also satisfied, our search for a minimax procedure is reduced to considering nonrandomized estimators of the form $c \sum X_i^2$ where the constant c is chosen to minimize the risk.

Note that Condition 2bp cannot be satisfied for the G used in the above example. If we had treated the example as a case of Example i′ so as to use Condition 2bp, we would have ended up searching through a much larger class of procedures unless we invoke some further principle such as that of sufficiency (in a manner similar to that of Sec. 4). We remark that Peisakoff indicated another method which could be used in some examples such as this one when one wants to use Condition 2bp: Let Q be a random variable independent of X and uniformly distributed on the component $O^+(n)$ of the identity of $O(n)$, and apply Condition 2bp to the G considered above on the sample space $R^{*n} \times O^+(n)$ of $X' = (X, Q)$. The disadvantage of using this technique, where it is possible to do so, is that in some examples further considerations may be required to reduce the class of invariant procedures to that which would have been obtained if Condition 2b′ had been used directly. Note that the technique used here is really related to that of Remark 7, which would give the desired result more directly here, but which would still be inferior to the direct use of Condition 2b′ which does not require the technique of Remark 7 in the present example.

(v) $\mathfrak{X}$ and G are as in Example iii, but with $n = 1$. $D = R^1$, the object being to estimate θ_1. The weight function is, e.g., a function of $(d - \theta_1) / \theta_2$, which we hereafter take to be the argument of W. There is one K_x, and if we try to verify Condition 2b′ we run into trouble. For example, take $x_x = 0$ so that M is the multiplicative group of reals (not normal in G) and $\mathfrak{X} = G/M$; M does not act trivially on D, so Condition 2b′ is not satisfied. If we consider this example as a case of Example i (i.e., let G be the smaller group used there), we obtain

for $\mathfrak{D}_I$ the class C of procedures δ for which $\delta(x, \Delta + x) = \delta(0, \Delta)$. If the conclusion of the theorem of Sec. 3 were valid for G = affine group, we could restrict ourselves to those members of C for which $\delta(x, \Delta) = \delta(ax, a\Delta)$ for all $a > 0$; putting $x = 0$, this means $\delta(0, \Delta) = \delta(0, a\Delta)$ for all $a > 0$; taking Δ to be the interval $(-1, 1)$, this means $\delta(0, 0) = 1$; noting the equation defining C, this means that there is only one invariant procedure δ^* under the affine group, the nonrandomized estimator $t(x) = x$. One would like to conclude that this estimator is minimax. If $\liminf_{t\to 0} W(t + X) \geqq W(X)$ w.p.1 when $(\theta_1, \theta_2) = (0, 1)$, an application of Fatou's lemma to the equation

$$r_\delta(\theta_1, \theta_2) = \int\int W\left(u + \frac{r}{\theta_2}\right) F_0(du)\delta(0, dr) \text{ if } \delta \in C$$

yields the fact that $\liminf_{\theta_2\to\infty} r_\delta(\theta_1, \theta_2) \geqq r_{\delta^*}(=$ constant) for $\delta \in C$, and the conclusion that δ^* is minimax is justified. However, Counterexample C below shows that without some such additional assumption as the one made here on W, *this conclusion is false*: δ^* need not be minimax and we can only conclude that there is a $\delta \in C$ which is minimax (or ϵ-minimax).

(vi) The univariate general linear hypothesis (GLH) is discussed in detail in many places. If γ is the parameter on which the power function of the usual F-test of specified size ϵ depends, it is easily proved (see, e.g., [5a]) that this test is uniformly most powerful invariant of size ϵ of the GLH $\gamma = 0$ (against $\gamma > 0$). There are several ways to apply the theorem of the next section to conclude, *e.g.*, that this test is most stringent of size ϵ (first proved by Hunt and Stein). One is to consider for fixed $\gamma_0 > 0$ the problem of testing $\gamma = 0$ against $\gamma = \gamma_0$, to note that G is transitive on ω and on $\mathfrak{F} - \omega$ in this case so that Assumption 3 is satisfied (as are the other assumptions), and thus to conclude that the above test is most stringent of size ϵ; since this is true for every $\gamma_0 > 0$, it follows that the test is most stringent for the original GLH. Another method (better than the above in other problems where such a property uniform in γ_0 may not hold) is to verify Assumption 3 directly for GLH; we can do this easily by applying the theory of [15] to the present case. Alternatively, (2.3) can be verified by considering, on the right side of (2.3), a ξ assigning probability one to the set consisting of one point in ω and one point at which the power function of the F-test differs most from the envelope power function.

COUNTEREXAMPLES. We now list briefly four counterexamples to the conclusion of the theorem of Sec. 3, only the third of which is new, in order to indicate that Assumptions 1, 2, 4, and 5 cannot be entirely dispensed with.

(A) In [6] and also in [7] are given examples which show that the conclusion of the theorem of Sec. 3 is false if (in terms of the present treatment) Assumption 4 is violated. We note here also that if, in the notation of p. 313 of [7], the weight function is altered to $f(s) = 1$ if s is an integer and $f(s) = \max(s, 0)$ otherwise, then there exist invariant procedures with finite risk ($=1$), but the conclusion of the theorem is still false; thus, we see that if in Condition 4c the condition that $W(y, z) \to c_z$ as $y \to -\infty$ and $\to\infty$ monotonically as $y \to \infty$ were dropped while

maintaining the condition that a $\delta_0 \in \mathfrak{D}_I$ with finite risk exists, Assumption 4 would not be implied.

(B) As Peisakoff has pointed out, the invariance theory applies to the general sequential case only if we restrict $\mathfrak{D}$ to consist of procedures which take at least a first observation with probability one. In Section 4 we shall discuss this in more detail (there are cases where this restriction of $\mathfrak{D}$ is not necessary); for the moment, we give an example to demonstrate that the conclusion of the theorem of the next section would not generally be true without such a restriction. Suppose we are limited to taking a single observation or else no observation on a random variable whose distribution depends only on a location parameter θ which we desire to estimate (see Example (i)), the loss from estimating θ incorrectly being bounded by 1 and the cost of experimentation being 2 or 0 depending on whether or not we take an observation. Any minimax procedure in $\mathfrak{D}$ must clearly take no observation with probability $\geq \frac{1}{2}$ (a similar remark applying for ϵ-minimax procedures); however, the only invariant procedures take a first observation with probability one (see Sec. 4 for further discussion). The difficulty here is that Assumption 1 is violated, since g_x must depend on the observation and thus, for a δ which requires no observations, the δ_g of (2.1) would require no observations but would depend on the observation, and would thus not be a legitimate decision function.

(C) As an example which shows that Assumption 2 cannot be entirely dispensed with, consider the setup of Example v with $F_0(x) = 0$ if $x < 0$ and $=1$ if $x \geq 0$, and let $W = 1$ if $d = \theta$ and $=0$ otherwise. This is essentially a game where one player says "don't you name the real number I name" and then names a real number, while the only affine-invariant procedure for the other player is, on hearing the number, to name the same number. The procedure δ^* of Example v is in fact uniformly worst and is clearly not minimax, while there exist many minimax procedures in the class C. This example can be made into one where all members of $\mathfrak{F}$ have densities with respect to a fixed σ-finite measure by restricting $\mathfrak{X}$, D, and G to the rationals (of course, this changes μ, and Condition 5b(2) is no longer applicable), and can be made more probabilistic by letting F_0 assign probability $\frac{1}{3}$ to each of the values -1, 0, 1; but the phenomenon persists. See Example v for an example of a condition which eliminates the phenomenon encountered in Counterexample C.

(D) Stein [17] has announced an example in testing hypotheses where all our assumptions except Assumption 5 are satisfied and where the conclusion of the theorem is false. This example shows that the real projective group and $GL(2)$ do not satisfy Assumption 5.

3. Proof of invariance theorem. We now use a modification of the method of proof used in [1] under Condition 2bp and Assumptions 4 and 5, in order to prove the following theorem (see also Remark 7 of Sec. 2):

THEOREM. *If G leaves the problem invariant and if Assumptions 1 to 5 are satisfied, then for any $\delta \in \mathfrak{D}$ and $\epsilon > 0$ there is a $\delta' \in \mathfrak{D}_I$ such that $r_{\delta'} \leq \epsilon + r_\delta$. In*

particular, if δ^* is minimax among procedures in $\mathfrak{D}_I$, then it is minimax among procedures in $\mathfrak{D}$.

PROOF. Our first step is to prove (3.5) below. Denote right invariant measure on G by μ^{-1}; i.e., $\mu^{-1}(E) = \mu(E^{-1})$. Fix b and $\delta \in \mathfrak{D}$ and let $\{G_n\}$ be a sequence satisfying Assumption 5, and define

$$H_{F,x}(g) = \int_D W^b(F, x, gr)\delta(g^{-1}x, dr). \tag{3.1}$$

Then, for $\gamma \in G$,

$$\begin{aligned} &\lim_{n\to\infty} \left| \int_{G_n^{-1}} [H_{F,x}(g^{-1}) - H_{F,x}(\gamma g^{-1})]\mu^{-1}(dg)/\mu(G_n) \right| \\ &\qquad = \lim_{n\to\infty} \left| \int_{G_n} [H_{F,x}(h) - H_{F,x}(\gamma h)]\mu(dh)/\mu(G_n) \right| \\ &\qquad \leqq \lim_{n\to\infty} 2b\mu(\gamma G_n \Delta G_n)/\mu(G_n) = 0, \end{aligned} \tag{3.2}$$

by Assumption 5. Using (3.2) with $\gamma = g_x$ and bounded convergence, we obtain, for any fixed $\xi \in S_b$,

$$\lim_{n\to\infty} \int_{\Gamma_b} \xi(dF) \int_{\mathfrak{X}} F(dx) \int_{G_n^{-1}} [H_{F,x}(g^{-1}) - H_{F,x}(g_x g^{-1})]\mu^{-1}(dg)/\mu(G_n) = 0. \tag{3.3}$$

It will simplify notation if we define the operation L by

$$L = \lim_{n\to\infty} \inf \int_{\Gamma_b} \xi(dF) \int_{\mathfrak{X}} F(dx) \int_{G_n^{-1}} \mu^{-1}(dg)[\mu(G_n)]^{-1} \int_D . \tag{3.4}$$

Using (3.3), a change of variables, and (2.1), we obtain

$$\begin{aligned} LW^b(F, x, g^{-1}r)\delta(gx, dr) &= LW^b(F, x, g_x g^{-1} r)\delta([g_x g^{-1}]^{-1}x, dr) \\ &= LW^b(F, x, u)\delta(gx^*, d_u g g_x^{-1} u) \\ &= LW^b(F, x, u)\delta_g(x, du). \end{aligned} \tag{3.5}$$

Let $\delta \in \mathfrak{D}$. Using the fact (Assumption 3) that S_b includes every measure giving probability one to a single element of $\Gamma_b \supset \{F_\beta\}$ and that gX has probability measure gF when X has measure F, we have for any fixed $\delta \in \mathfrak{D}$,

$$\begin{aligned} \bar{r}_\delta &= \sup_{F \in \bar{\Omega}} \sup_b \int_{\mathfrak{X}} F(dx) \int_D W^b(F, x, r)\delta(x, dr) \\ &= \sup_b \sup_{\xi \in \bar{S}_b} \sup_{g \in G} \int_{\Gamma_b} \xi(dF) \int_{\mathfrak{X}} F(dx) \int_D W^b(gF, gx, r)\delta(gx, dr) \\ &\geqq \sup_b \sup_{\xi \in \bar{S}_b} \liminf_{n\to\infty} \int_{G_n^{-1}} \mu^{-1}(dg)[\mu(G_n)]^{-1} \\ &\qquad \cdot \int_{\Gamma_b} \xi(dF) \int_{\mathfrak{X}} F(dx) \int_D W^b(gF, gx, r)\delta(gx, dr), \end{aligned} \tag{3.6}$$

where the inequality follows from the fact that an average is no greater than a supremum. Using Fubini's theorem ($\mu^{-1}(dg)$ on G_n^{-1} and $\xi(dF)F(dx)$ on $\Gamma_b \times \mathfrak{X}$ are both finite) and the invariance of W (i.e., $W^b(gF, gx, r) = W^b(F, x, g^{-1}r)$), we see that the last member of (3.6) is equal to the supremum with respect to b and ξ of the first member of (3.5). On the other hand, again using Fubini's theorem, the supremum with respect to b and ξ of the last member of (3.5) is equal to

$$\sup_b \sup_{\xi \in \mathfrak{F}_b} \liminf_{n\to\infty} \int_{G_n^{-1}} \mu^{-1}(dg) r^b_{\delta_g}(\xi)/\mu(G_n) \geqq \sup_b \sup_{\xi \in \mathfrak{F}_b} \inf_{\delta \in \mathfrak{D}_I} r^b_\delta(\xi), \tag{3.7}$$

the inequality following from the fact that an average is no less than an infimum and that $\delta_g \in \mathfrak{D}_I$ for $g \in G$ (see Remark 3). Using first Assumption 3 and the fact that $\sup_{\xi \in \mathfrak{F}_b} r^b_\delta(\xi) = \bar{r}^b_\delta$ if $\delta \in \mathfrak{D}_I$, and then using Assumption 4, we see that the right side of (3.7) is equal to

$$\sup_b \inf_{\delta \in \mathfrak{D}_I} \bar{r}^b_\delta = \inf_{\delta \in \mathfrak{D}_I} \bar{r}_\delta . \tag{3.8}$$

Thus, for each $\delta \in \mathfrak{D}$, the first member of (3.6) is no less than the last member of (3.8), proving the theorem.

The above theorem does not, of course, treat the question of whether or not a minimax procedure exists, i.e., whether $\inf_{\delta \in \mathfrak{D}} \bar{r}_\delta$ is attained. Conditions for this may be found, e.g., in [15] and [16]; the same conditions will usually apply for both $\mathfrak{D}$ and $\mathfrak{D}_I$, so that the conclusion of our theorem can be strengthened by the additional remark that a minimax procedure exists in $\mathfrak{D}_I$ if one exists in $\mathfrak{D}$. Various conditions for the attainment of $\inf_\delta \bar{r}_\delta$ are also given in [1] and [4] (see [5a]). Of course, for suitably simple W one can often write down an explicit formula for a minimax invariant procedure in the manner discussed under Condition NR of Sec. 2; for example, by now this formula is well known in the case studied in [6].

It is of interest to note an observation of Peisakoff to the effect that his proof (under Condition 2bp) will go through in many cases where the elements of $\mathfrak{F}$ are not *all* the distributions gF_0 for $g \in G$, but only a suitably large subset of these: e.g., in Example i of Sec. 2, the restriction $\theta \geqq 0$ might be imposed. This extension can also be carried out under our assumptions in certain cases where the restricted class of elements g for which $gF_\delta \in \mathfrak{F}$ is not compact.

4. The sequential case. Our setup in this section is that of Secs. 2 and 3 with certain interpretations. For simplicity our description is specialized to handle the examples stated at the end of this section, although a more general setup is obvious. The space $\mathfrak{X}$ is a product space $\mathfrak{X}_1 \times \mathfrak{X}_2 \times \cdots$ with denumerably many factors or a trivial modification of such a space as in Example ii or iv of Sec. 2, and we write a point of $\mathfrak{X}$ as $x = (x_1, x_2, \cdots)$ and the random variable X as $(X_1, X_2, \cdots)$. In the examples we treat, the $\mathfrak{X}_i$ will be copies of the same Euclidean space and the X_i will be independent and identically distributed according to each $F \in \mathfrak{F}$. The group G will act componentwise on $\mathfrak{X}$, so we may write $gx = (gx_1, gx_2, \cdots)$. The space D will be a product space $D_1 \times E$ where

the "terminal decision space" D_1 has the role the space D had in fixed sample-size problems and $B_D = B_{D_1} \times B_E$ where (B_{D_1}, D_1) is the Borel sets on a subset of a Euclidean space and B_E contains at least the countable subsets of E. The "experimental decision space" E consists of all ordered k-tuples of (not necessarily distinct) positive integers for $k = 0, 1, 2, \cdots$, as well as infinite sequences of positive integers; we represent an element of E by $e_k = (a_1, a_2, \cdots, a_k)$, such a k-tuple representing an experiment carried out in k stages, the ith of which consisted of a_i "observations," namely, on $X_{s_{i-1}+1}, \cdots, X_{s_i}$, where we write $s(e)$ for the sum of the integers in e and $s_k = s((a_1, \cdots, a_k))$; e_0 represents the taking of no observations, and we write $e_\infty = (a_1, a_2, \cdots)$ for an e where experimentation never ceases. The group G acts trivially on E, so that we may write $gd = g(d_1, e) = (gd_1, e)$ in the sequel. The weight function W can depend on F, d_1, and e; for simplicity of exposition, in this section the weight function W will be a sum of two non-negative parts:

$$W(F, x, (d_1, e)) = W_1(F, d_1) + W_2(e), \tag{4.1}$$

although the more general form $W(F, d_1, e)$ can be treated in similar fashion. Thus, W_1 takes the place of the W of the fixed sample-size case and must satisfy the invariance condition $W_1(gF, gd_1) = W_1(F, d_1)$ for all F, d_1 and g. The cost of experimentation $W_2(e_k)$ we assume to be non-negative and finite if $k < \infty$ and infinite if $k = \infty$ (the cases where $W_2(e_k)$ is permitted to be infinite for $k < \infty$ in some treatments of decision theory to reflect upper limits on sampling will be covered by restricting $\mathfrak{D}$ as indicated in Remark 8 below), and we assume the existence of a finite number q and a real nondecreasing function h tending to infinity with its non-negative argument and such that, for all $k < \infty$ and e,

$$\begin{aligned} W_2((a_1, \cdots, a_k, 1)) \quad &- W_2((a_1, \cdots, a_k)) < q, \\ W_2(e) \quad &> h(s(e)), \\ W_2((a_1, \cdots, a_k, a_{k+1})) &\geqq W_2((a_1, \cdots, a_k)); \end{aligned} \tag{4.2}$$

in other words, the cost of taking *one* additional observation at any stage is bounded, for any finite number M only finitely many different e's cost less than M, and additional observations always have non-negative cost. One often imposes on W_2 practical restrictions such as $W_2((a_1 + a_2)) \leqq W((a_1, a_2))$, but this is inessential for our considerations. Typical specializations of W_2 often encountered in practice are $W_2((a_1, \cdots, a_k)) = \sum_1^k W_2((a_i))$ or $= W_2((\sum_1^k a_i))$ the latter case with $W_2((t)) = ct$ being especially important.

Denote by B_n the Borel field of members of B_X which are cylinder sets with base in $\mathfrak{X}_1 \times \cdots \times \mathfrak{X}_n$; i.e., a B_n-measurable real function of x is one which depends on x only through $(x_1, \cdots, x_n)$, the only B_0-measurable functions being constants. We denote by $\mathfrak{D}^0$ the class of all sequential decision functions δ, i.e., functions δ on $\mathfrak{X} \times B_D$ which are probability measures on D for each x (see also the discussion of the paragraph containing (4.3) below for interpretation) where, in addition to the measurability requirements of Section 2, each $\delta \in \mathfrak{D}^0$ is assumed

to satisfy the restriction that if $e = (a_1, \cdots, a_k)$ with $s(e) = r$ and if $Q_{e,a}$ is the set of all elements e_∞ or e_k of E of the form $e_\infty = (a_1, \cdots, a_k, a, \cdots)$ or $e_j = (a_1, \cdots, a_k, a, a_{k+2}, \cdots, a_j)$ for all $j \geqq k+1$ and all $a_{k+2}, \cdots$, then $\delta(x, \Delta_1 \times e)$ (for each $\Delta_1 \in B_{D_1}$) and $\delta(x, D_1 \times Q_{e,a})$ are B_r-measurable in x; that is, the decision to stop taking observations or to take a particular number of observations at the next stage depends only on observations which have already been taken. Let $\mathfrak{D}^i$ denote the class of all δ in $\mathfrak{D}^0$ for which $\delta(x, D_1 \times e) = 0$ whenever $s(e) < i$; i.e., which for each x observe at least $x_1, \cdots x_i$ w.p.1. For $i \geqq 0$, let $\mathfrak{D}^i_I$ denote the invariant procedures in $\mathfrak{D}^i$; of course, δ is invariant if $\delta(gx, g\Delta_1 \times e) = \delta(x, \Delta_1 \times e)$ for all g, x, Δ_1, e. We have already seen in Counter-example B of Sec. 2 that the theorem of Sec. 3 will not generally be true if $\mathfrak{D} = \mathfrak{D}^0$ because not all of the δ_ε of Assumption 1 will be decision functions. Of course, if G were compact we could use the method of [4] directly as outlined in Sec. 1, without any difficulty. For the examples treated at the end of this section it will suffice to take $\mathfrak{D} = \mathfrak{D}^1$ or $\mathfrak{D}^2$. (The sequential considerations of [1] consist of briefly pointing out an example of the sequential setup of $\mathfrak{D}$ and the necessity of not taking $\mathfrak{D} = \mathfrak{D}^0$.)

The question arises, how much do we lose by restricting $\mathfrak{D}$ to be $\mathfrak{D}^1$ or $\mathfrak{D}^2$ rather than $\mathfrak{D}^0$? The answer will usually be easy to verify. For example, suppose D_1, G, $\mathfrak{F}$, and the X_i are as in Example i (or i') of Sec. 2 (Examples vii to x of the present section) and that W_1, which we may think of as a function of $\theta - d_1$, tends to its supremum w (say) when its argument tends to ∞ (or, similarly, $-\infty$). Then any procedure δ which requires 0 observations w.p.1 clearly has $\bar{r}_\delta = w$. Since any member of $\mathfrak{D}^0$ can be written as a probability mixture of a procedure in $\mathfrak{D}^1$ and a procedure which requires 0 observations w.p.1, it is evident that either every procedure requiring 0 observations w.p.1 is minimax, or else there is a $\delta \in \mathfrak{D}^1$ which is minimax. Which of these is the case will be easy to verify in most practical examples. In particular, if $w = \infty$, the second is always the case.

The function δ as given above is (with a different notation) the function p defined in Eq. (1.3) of [15]; $\delta(x, \Delta_1 \times e)$ is the probability, when δ is used and $X = x$, that the experiment will terminate with experimental decision e and terminal decision an element of the subset Δ_1 of D_1. The usual representation of a sequential decision function is obtained by letting $\bar{D}$ be the union of D_1 with the space L of positive integers and writing, for each element e of E and subset $\bar{\Delta}$ of $\bar{D}$,

$$\delta(\bar{\Delta} \mid x, e) = \frac{\delta(x, \bar{\Delta}_1 \times e) + \delta(x, D_1 \times Q'_e)}{\delta(x, D_1 \times Q_e)}, \tag{4.3}$$

where Q_e is the set of all elements of the form $e_\infty = (a_1, \cdots, a_k, \cdots)$ or $e_j = (a_1, \cdots, a_k, \cdots, a_j)$ of E for all $j \geqq k$, when $e = (a_1, \cdots, a_k)$ (thus, Q_e is the union of all $Q_{e,a}$ for $a > 0$, together with e), while Q'_e is the union over $a \in \bar{\Delta} \cap L$ of the sets $Q_{e,a}$, and we let $\bar{\Delta}_1 = \bar{\Delta} \cap D_1$. If the denominator of the right side of (4.3) is 0, define $\delta(\bar{\Delta} \mid x, e) = 1$ or 0 according to whether or not $1 \in \bar{\Delta} \cap L$; the

definition in this case is only for definiteness and could be made in many other ways. The left side of (4.3) represents the conditional probability, when δ is used and given that $X = x$ and that the experiment has already proceeded (if $e = (a_1, \cdots, a_k)$) through k stages of experimentation as represented by e, that a terminal decision in $\bar{\Delta}_1$ is made or that the next stage of the experiment consists of a number of observations in $\bar{\Delta}_2 = \bar{\Delta} \cap L$. Clearly, $\delta(\bar{\Delta} \mid x, e)$ is B_r-measurable in x if $s(e) = r$, and the functions $\delta(\bar{\Delta} \mid x, e)$ on $B_{\mathfrak{D}} \times \mathfrak{X} \times E$ satisfying obvious restrictions are in 1-to-1 correspondence with the functions $\delta(x, \Delta)$ on $\mathfrak{X} \times B_D$ as described originally ($B_{\mathfrak{D}}$ consists of every union of a set in B_{D_1} and a set in B_E). Moreover, in terms of our later description, δ is invariant if $\delta(\bar{\Delta} \mid x, e) = \delta(g\bar{\Delta} \mid gx, e)$, where $g\bar{\Delta} = g(\bar{\Delta}_1 \cup \bar{\Delta}_2) = (g\Delta_1) \cup \Delta_2$. We shall use this representation of $\mathfrak{D}_I^*$ below.

The problems we are going to consider are ones in which the difficulty encountered in Counterexample B can be avoided as indicated above, and in which there is a very simple sufficient sequence $\{T_i\}$ of functions on $\mathfrak{X}$, T_i being B_i-measurable (the range space of T_i is immaterial). If one does not employ the principle of sufficiency in the manner of this section the theorem of Sec. 3 will only yield the dependence of the stopping rule on $x_\alpha (= (x_2 - x_1, \cdots, x_n - x_1)$ after n observations in Example vii, for example), nothing like the result we obtain. Specifically, we assume (see Example xv for further remarks)

ASSUMPTION 6. *For some positive integer m, Assumptions 1 and 2 are satisfied for $\mathfrak{D} = \mathfrak{D}^m$ with g_x a B_m-measurable function of x. There exists a sequence $\{T_i\}$ of functions with T_i a B_i-measurable sufficient statistic for $[(X_1, \cdots, X_i), \mathfrak{F}]$, such that there exist conditional probability* d.f.'s

$$F_r(y_1, \cdots, y_r \mid t_r) = P\{g_x^{-1}(X_1, \cdots, X_r) \leqq (y_1, \cdots, y_r) \mid T_r(X) = t_r\}$$

for $r \geqq m$ with the property

(4.4) $\qquad F_r(y_1, \cdots, y_r \mid t_r)$ *does not depend on* t_r.

It will aid understanding to consider an example at this point, Example vii of this section. The X_i are normal with unknown mean and known variance, and $\mathfrak{X}_i = G = D_1 = R^1$. We also identify $\mathfrak{F}$ with R^1 in obvious fashion. We can let $n = 1$ and $g_x u = u + x_1$ for u in $\mathfrak{X}_i$ or D_1, and identify the indices α with sequences $x_\alpha = g_x^{-1}x = (0, x_2 - x_1, x_3 - x_1, \cdots)$. Let $T_i = \sum_{j=1}^{i} X_j$. Since $g_x^{-1}X_1 = 0$ and $g_x^{-1}(X_2, \cdots, X_r) = (X_2 - X_1, \cdots, X_r - X_1)$, the distribution of $g_x^{-1}(X_1, \cdots, X_i)$ given that $T_i(X) = t_i$ is multivariate normal with means and covariances independent of t_i, so that Assumption 6 is satisfied. Similarly, in Example xi with G the affine group and $\mathfrak{X}_i = R^1$, we put $g_x^{-1}x_i = (x_i - x_1)/(x_2 - x_1)$, etc.

Assumption 6 is related to a property cited in [5a] as being proved in [4] in certain regular cases, to the effect that we lose nothing in the validity of the theorem of Sec. 3 for problems considered in [4] if we first use the principle of sufficiency and then apply the invariance principle to the space of a correctly chosen sufficient statistic. Assumption 6 also includes an additional strong

property in that (4.4) is obviously not implied by this result of [4] (see also Example xv below). This assumption is easily verified in Examples vii to xiv.

Denote by $Q(s)$ the infimum of $\bar{r}_\delta - W_2(e_1^{(s)})$ over all δ with $\delta(x, D_1 \times e_1^{(s)}) = 1$, where $e_1^{(s)} = (s)$, and by $Q_I(s)$ the infimum when δ is also restricted to be invariant; thus, $Q(s)$ and $Q_I(s)$ are the values of $\inf_\delta \bar{r}_\delta$ over all δ or all invariant δ for the fixed sample-size problem with sample size s when the weight function is W_1. We assume

ASSUMPTION 7. *Either $\bar{r}_\delta = \infty$ for all $\delta \in \mathfrak{D}^m$ or else there is an integer m' with $Q(m') < \infty$.*

This assumption is easy to verify in practical cases for the examples considered in this section, where one will usually know $Q_I(j) < \infty$ for some j. The assumption can be shown, in fact, to be implied by our other assumptions under mild regularity conditions, although for the sake of brevity we forego such considerations here.

The main remaining difficulty in applying the theorem of Sec. 2 to the present problem is the verification of Assumption 4, which would usually be difficult to verify directly in sequential problems. Our form of the theorem which follows reduces this verification to the much simpler nonsequential one of Sec. 2.

THEOREM. *If G leaves the problem invariant and Assumptions* 3, NR, 5, 6, 7, (4.1), *and* (4.2), *as well as Assumption* 4 *for W_1 in each fixed sample-size problem with sample size $\geqq m$, are satisfied, and if $\mathfrak{D} = \mathfrak{D}^m$, then for each $\epsilon > 0$ there exists a fixed sample-size invariant procedure δ^* (the sample perhaps being taken according to some grouping) with $\bar{r}_{\delta^*} \leqq \epsilon + \inf_{\delta \in \mathfrak{D}} \bar{r}_\delta$. Thus, if $Q_I(s(e)) + W_2(e)$ is minimized over $s(e) \geqq m$ by $e = e'$ and if δ^* is a minimax invariant procedure for the fixed sample-size problem with sample size $s(e')$ ignoring W_2, then a minimax procedure for the sequential problem is to take $s(e')$ observations according to the grouping e' (which minimizes $W_2(e)$ over e satisfying $s(e) = s(e')$) and then to use δ^*.*

REMARK 8. Before proving the theorem we remark that the first paragraph of the proof below can easily be altered to handle the case where $\mathfrak{D}$ is further restricted in some way such as bounding k or the a_i or $s(e_k)$ in $e_k = (a_1, \cdots, a_k)$, etc. We have already noted the fact that it will usually be easy to verify whether a minimax procedure of $\mathfrak{D}^m$ or a more trivial procedure is minimax in $\mathfrak{D}^0$. We also note that one can think of G as acting on T_r for $r \geqq m$ in the examples treated by us, so that the conclusion of the theorem could be phrased in terms of invariant functions of T_r.

PROOF OF THEOREM. We may assume $\rho = \inf_{\delta \in \mathfrak{D}} \bar{r}_\delta < \infty$, the theorem being trivial otherwise. By Assumption 7 there is an m' and a procedure δ^0 with $\delta^0(x, D_1 \times e_1^{(m')}) = 1$ and $\bar{r}_{\delta^0} - W_2(e_1^{(m')}) = C < \infty$. Since the X_i are independent and identically distributed we can clearly assume $m' \geqq m$. Let ϵ be a positive number. The second line of (4.2) implies the existence of a number $N' > m$ such that any procedure $\delta \in \mathfrak{D}^m$ with $\bar{r}_\delta < \rho + \epsilon$ must require fewer than N' observations with probability $> 1 - \epsilon$ for all $F \in \mathfrak{F}$. For any such δ define the procedure δ' as one which proceeds like δ except that whenever ex-

perimentation has reached a stage e (including e_0) where $s(e) < N' \leqq s(e) + t$ for some t with $\delta(x, e_1^{(t)} \mid e) > 0$, δ' assigns the probability $\delta(x, e_1^{(t)} \mid e)$ which δ assigned to the taking of t observations at the next stage (there may of course be several such t) to the taking of exactly m' additional observations one-by-one and, if these observations are taken, uses δ^0 on these last m' observations to reach a terminal decision. Since the X_i are independent and identically distributed, by the first and last lines of (4.2) we clearly have $\bar{r}_{\delta'} < \bar{r}_\delta + \epsilon(C + qm')$ and $\delta' \in \mathfrak{D}^m$. Since $\epsilon > 0$ is arbitrary, we conclude that our theorem will be proved if we prove it for the case where $\mathfrak{D}$ is restricted to the class $\mathfrak{D}^{m,N}$ of procedures in $\mathfrak{D}^m$ for which $\delta(x, D_1 \times E_N) = 1$, where N is a fixed integer and E_k is the set of e for which $s(e) < k$. We hereafter assume $\mathfrak{D} = \mathfrak{D}^{m,N}$.

In order to apply the theorem of Sec. 3 to the present case, it remains to verify Assumption 4 when $\mathfrak{D} = \mathfrak{D}^{m,N}$. Let $Q_I^b(s)$ be the value of $Q_I(s)$ when W_1 is replaced by W_1^b. By Assumption 4 in the fixed sample-size case, we have

$$\lim_{b\to\infty} [W_2(e) + Q_I^b(s(e))] = W_2(e) + Q_I(s(e))$$

for each fixed e with $s(e) \geqq m$. Since there are only finitely many e with $m \leqq s(e) < N$, we obtain, for $\mathfrak{D} = \mathfrak{D}^{m,N}$,

$$\begin{aligned}\lim_{b\to\infty} \inf_{\delta \in \mathfrak{D}_I} \bar{r}_\delta^b &= \lim_{b\to\infty} \min_{m \leqq s(e) < N} [W_2(e) + Q_I^b(s(e))] \\ &= \min_{m \leqq s(e) < N} [W_2(e) + Q_I(s(e))] = \inf_{\delta \in \mathfrak{D}_I} \bar{r}_\delta,\end{aligned}$$

which is Assumption 4 for the present problem.

Applying, then, the theorem of Sec. 3, we obtain for any $\delta \in \mathfrak{D}^{m,N}$ and $\epsilon > 0$ an invariant procedure δ' with $\bar{r}_{\delta'} \leqq \bar{r}_\delta + \epsilon$. Since δ' is invariant, we have

$$\begin{aligned}\delta'(x, \Delta \mid e) &= \delta'(g_x^{-1}x, g_x^{-1}\Delta \mid e) \\ &= \delta'(g_x^{-1}x, g_x^{-1}\Delta_1 \mid e) + \delta'(g_x^{-1}x, \Delta \cap L \mid e).\end{aligned}$$

Define the procedure δ'' by

$$\begin{aligned}\delta''(x, \Delta \mid e) = {} & E\{\delta'(g_X^{-1}X, g_X^{-1}\Delta_1 \mid e) \mid T_{s(e)} = T_{s(e)}(x)\} \\ & + \int \delta'(y, \Delta \cap L \mid e) F_{s(e)}(dy_1, \cdots, dy_{s(e)} \mid T_{s(e)}(x)).\end{aligned} \tag{4.5}$$

Since B_{D_1} = Borel sets on a Euclidean set, this defines a decision function for some version of the conditional expected value (see, e.g., [18]). Clearly $\delta''(x, D_1 \times E_k) = 0$ for $k = m$ and $= 1$ for $k = N$, so $\delta'' \in \mathfrak{D}^{m,N}$. Since $\{T_i\}$ is sufficient, $r_{\delta'} = r_{\delta''}$. But for each e and $\Delta_2 \subset L$, Assumption 6 implies that $\delta''(x, \Delta_2 \mid e)$ is a constant. Hence δ'' can be considered to be a member of the class ϕ of probability mixtures of fixed sample-size procedures of sample-sizes $s(e)$ with $m \leqq s(e) < N$, where the sample may be taken according to some grouping e (independent of X). It is easy to see that, under our assumptions, the result of the previous paragraph remains true if $\mathfrak{D}^{m,N}$ is replaced by ϕ and that Assumptions 1, 2, 3, and 5 remain satisfied; thus, the theorem of Sec. 3 is valid for $\mathfrak{D} = \phi$,

so that there is a $\delta^* \varepsilon \phi_I \subset \mathfrak{D}_I^{m,N}$ with $\bar{r}_{\delta^*} \leqq \bar{r}_{\delta''} + \epsilon \leqq \bar{r}_\delta + 2\epsilon$. This completes the proof of the theorem, since condition NR implies the constancy of r_{δ^*} and hence the existence of a fixed sample-size $\delta^{**} \varepsilon \phi_I$ with $r_{\delta^{**}} \leqq r_{\delta^*}$.

We note that δ'' in the preceding paragraph can be proved invariant in our examples, for an appropriate version of the first term on the right in (4.5), but the proof as given seems just as short. The lack of dependence of W on x in (4.1) is of course used in invoking sufficiency.

EXAMPLES. We shall use the following notation in our examples, where z and θ_1 are real and θ_2 and γ are positive:

$$f_1(z; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi}\theta_2} e^{-(z-\theta_1)^2/2\theta_2^2},$$

$$f_2(z; \theta_1, \theta_2) = \begin{cases} 1/2\theta_2 & \text{if } |z - \theta_1| < \theta_2 \\ 0 & \text{otherwise,} \end{cases}$$

$$f_{3\gamma}(z; \theta_1, \theta_2) = \begin{cases} (z - \theta_1)^{\gamma-1} e^{-(z-\theta_1)/\theta_2}/\theta_2^\gamma \Gamma(\gamma) & \text{if } z > \theta_1 \\ 0 & \text{otherwise.} \end{cases}$$

In all the examples except xiv, the X_i will be independent real random variables whose common Lebesgue density will be assumed to be in some class of the above densities, which class we identify with $\mathfrak{F}$.

(vii) $\mathfrak{F}$ consists of the densities f_1 for $-\infty < \theta_1 < \infty$ with θ_2 assumed known and θ_1 to be estimated and hence $G = D_1 = \mathfrak{F}$ = additive group of R^1, W_1 being a function of $\theta_1 - d_1$. Note that in most practical examples Assumption 4 can be verified by applying Condition 4a or 4b or 4c, and the question of whether to use a procedure requiring no observations or one in $\mathfrak{D}^1$ will be easy to settle. Of course, we can take $T_i = \sum_{j=1}^{i} X_j$, $m = 1$, and $g_x u = u + x_1$, as previously mentioned. Thus, the conclusion of the theorem will be satisfied for most W_1 and W_2 encountered in practice. Of course, $Q_I(s)$ is easily computed in this case to be given by

$$Q_I(s) = \inf_h \int_{-\infty}^{\infty} W_1(h + u) f_1(u; 0, \theta_2 s^{-\frac{1}{2}})\, du; \tag{4.6}$$

and, if h_s achieves the minimum, a nonrandomized sequential minimax estimator will be given by taking $s(e')$ observations according to the grouping e' described in the statement of the theorem and then estimating θ_1 by $s(e')^{-1} T_{s(e')} + h_{s(e')}$.

(vii+) We mention several extensions of Example vii: (1) The form of the minimax (or an analogous ϵ-minimax) estimator above depends on θ_2 in such a way that if it were only known that θ_2 belonged to some set B (not necessarily the set of all positive numbers) and if W_1 were a function of $(d_1 - \theta_1)\theta_2^{-1}$ instead of $(d - \theta_1)$, then the estimator of the previous problem vii for the case $\theta_2 = 1$ would be minimax (or ϵ-minimax) here. (2) A second extension is to note that, for the original setup of Example vii, if W_1 is symmetric we can also apply the group of reflections as in Example iii of Sec. 2. If in addition W_1 is nondecreasing

in $|\theta_1 - d_1|$, we obtain the sample mean ($h_s = 0$ in vii) as minimax estimator, a result first obtained in [8], with a special case in [9]. Note that the question of whether a procedure in $\mathfrak{D}^1$ or one requiring no observations is minimax is trivial in this case.

(viii) Same as vii, except that the possible distributions are the $f_{2\gamma}$ with $\gamma = 1$ and θ_2 known and θ_1 unknown with $-\infty < \theta_1 < \infty$. In this case $T_i = \min(X_1, \cdots, X_i)$ and the considerations and conclusions are as in vii with f_1 replaced by $f_{21}(u; 0, \theta_2 s^{-1})$ in (4.6), the minimax estimator being $T_{s(s)} + h_{s(s)}$. A very special case of this was obtained tediously in [11].

(ix) $\mathfrak{F}$ consists of the densities f_2 with θ_1 known and θ_2 unknown, $0 < \theta_2 < \infty$. Here θ_2 is to be estimated, so $D_1 = \mathfrak{F} = G^{(1)} =$ multiplicative group of positive reals. The weight function is a function only of θ_2/d_1. We can either take $G = G^{(1)}$ or can think of θ_1 is being 0 and let $G =$ direct product of $G^{(1)}$ and $G^{(2)}$ where $G^{(2)}$ contains the identity and an element which multiplies X_i by -1 and leaves $\mathfrak{F}$ and D fixed. We have

$$m = 1 \quad \text{and} \quad T_i = \max(|X_1 - \theta_1|, \cdots, |X_i - \theta_1|)$$

and $g_x^{-1}u = u/(x_1 - \theta_1)$ if $G = G^{(1)}$, with an obvious modification if we let $G = G^{(1)} \times G^{(2)}$. In either case, Assumption 6 is satisfied. Of course, this problem is really the same as that of estimating θ_2 when X_i has density $1/\theta_2$ for $0 < x_i < \theta_2$ and 0 otherwise (put $X_1' = |X_i - \theta_1|$ above), and in this form the problem may be reduced to that of viii by a logarithmic transformation as in Example i′. The form of the analogue of (4.6) and of the minimax procedure are obvious. The special case $W_2 = (\theta_2 - d_1)^2/\theta_2^2$ was considered in [11]; Condition 4c is satisfied there.

(x) $\mathfrak{F}$ consists of the densities $f_{2\gamma}$ where γ and θ_1 are known and θ_2 is unknown, $0 < \theta_2 < \infty$. This is a scale parameter problem with $G =$ the $G^{(1)}$ of ix, and we need only remark that the theorem applies with $T_i = \sum_{j=1}^{i} X_j$, the analogues of (4.6) and the form of the minimax procedure being obvious. This problem was treated for a particular γ and weight function in [10] and for a special class of weight functions in [12].

(x′) If X_i has symmetric density about known θ_1, the density of $|X_i - \theta_1|$ being that of Example x, the same considerations apply, using also the $G^{(2)}$ of ix. Similarly, the problem of estimating θ_2 when f_1 is the density and θ_1 is known can obviously be reduced to that of Example x.

The next three examples are similar in that, in each, there is both an unknown location parameter θ_1 and also an unknown scale parameter θ_2 with $-\infty < \theta_1 < \infty$ and $0 < \theta_2 < \infty$. In each case $m = 2$, G is the real affine group (see Example ii), and $g_x^{-1}x_i = (x_i - x_1)/(x_2 - x_1)$. There are three main types of problems in each example: (1) estimation of both θ_1 and θ_2, so that $D_1 = G$, $d_1 = (d_{11}, d_{12})$, W_1 is a function of $(\theta_1 - d_{11})/\theta_2$ and d_{12}/θ_2, and

$$g_x^{-1} d_1 = ((d_{11} - x_1)/(x_2 - x_1), d_{12}/(x_2 - x_1));$$

(2) estimation of θ_1, where $D_1 = R_1$, W_1 is a function of $(\theta_1 - d_1)/\theta_2$, $g_x^{-1} d_1 = (d_1 - x_1)/(x_2 - x_1)$; (3) estimation of θ_2, where $D_1 =$ positive reals, W_1 is a func-

tion of d_1/θ_2, $g_x^{-1} d_1 = d_1/(x_2 - x_1)$; of course, (2) and (3) can really be considered as special cases of (1) where W_1 only depends on one of its two arguments. For each type and example it is simple to write down an analogue to (4.6) and the corresponding form of a minimax procedure. In each case the conditions of the theorem are easily verified for many commonly used W, and the verification of whether one should use a procedure in $\mathfrak{D}^2$ or one requiring one or no observations is also easy. The use of the Bayes method in these examples would of course be much more complicated than that in [8], [11], and [12].

(xi) $\mathfrak{F}$ consists of all densities f_1. Putting $\bar{X}^{(i)} = i^{-1} \sum_{j=1}^{i} X_j$, we have $T_i = (\bar{X}^{(1)}, \sum_{j=1}^{i} (X_j - \bar{X}^{(i)})^2)$ for $i \geqq 2$. Note that the problem of estimating θ_2, even for the appropriate weight function, cannot be obtained by the method of [10] without some modification, because of the nature of the Cramér-Rao bound.

(xii) $\mathfrak{F}$ consists of all densities f_2. Putting $U_i = \min(X_1, \cdots, X_i)$ and $V_i = \max(X_1, \cdots, X_i)$, we can take $T_i = (U_i, V_i)$ or $((U_i + V_i)/2, (V_i - U_i))$, for $i \geqq 2$. (The second form of T_i here and in the next example are pertinent to remarks made below in Example xv.)

(xiii) $\mathfrak{F}$ consists of all densities f_{31} (i.e., γ is known to be 1), and in the notation of the previous two examples we can take $T_i = (U_i, \bar{X}^{(i)})$ or $(U_i, \bar{X}^{(i)} - U_i)$ for $i \geqq 2$.

(xiv) As an example of a multivariate nature, suppose $\mathfrak{X}_i = R^J$ for some positive integer J, the X_i again being independent and identically distributed. Here $X_i = (X_{i1}, \cdots, X_{iJ})$, and we assume X_i has a multivariate normal distribution with the identity convariance matrix and unknown mean $\theta = (\theta_1, \cdots, \theta_J) \varepsilon R^J$. The problem is to estimate θ, so that $D_1 = \mathfrak{F} = G =$ additive group of R^J and W_1 is a function of the difference between the vectors d_1 and θ_1. Taking $m = 1$ and $T_i = \bar{X}^{(i)}$ and $g_x^{-1} u = u - x_1$ for $u \varepsilon R^J$, the theorem is applicable for many common weight functions. (Examples viii to xiii have similar multivariate analogues.)

(xiv+) We can extend Example xiv in the manner of vii+. In particular, if W_1 is an increasing function of the usual Euclidean *distance* between d_1 and θ, it is easy to see that $\bar{X}^{(a(a'))}$ is a minimax sequential estimator. The orthogonal group also leaves the problem invariant in this case, but this fact need not be used in obtaining the above form of the minimax estimator, it sufficing to apply a result of [19]. It is interesting to note that it is shown in [20] that, when W_1 is the squared length of the distance and $J > 2$, this estimator is not admissible.

(xv) As an example which illustrates the fact that the method of this section yields little if no T_i satisfy Assumption 6, consider the problem of estimating θ_1 when the X_i have density f_2 and θ_2 is known. This problem is considered for certain W in [15] and [21], and the minimax procedures obtained there are not fixed sample-size. As in Example xii, (U_i, V_i) is a minimal sufficient statistic. Assumption 6 cannot be satisfied for any sufficient T_i. The application of our method in this example would yield the form of the estimator obtained in [15] and [21], but would only yield the fact that the minimax stopping rule depends on $U_i - V_i$ at the ith stage; the stationary form of the minimax stopping rule seems

to depend strongly on the particular nature of f_2. It will be noted that the previous examples differ from this one in that in the former, but not in the latter, *there is a natural version of T_i for $i \geqq m$ whose range is G and such that the problem in terms of the T_i is left invariant by the natural operation of G on the range of T_i.* This is the essence of the examples where the method of this section yields the conclusion of the theorem, although we have seen that G may be modified somewhat from what this statement indicates (see Examples vii+, ix, x′, and xiv+) to the case where the *range of T_i is a subgroup or homogeneous space of G*. We may add that, in most sequential *testing* problems, the invariance principle yields little, for reasons similar to those present in Example xv.

REMARK 9. We end this section with a remark about other versions of the statistical problem, such as that of minimaxing the W_1 component of the risk subject to a bound on the W_2 component or vice versa. This includes such problems as the problem of finding optimum sequential estimators of bounded relative error of the scale parameters in Examples ix to xiv (in [7] there is some discussion of this problem but our results are not obtained) and that of obtaining optimum sequential interval estimators of prescribed length and confidence coefficient for the location parameters in Examples vii and viii. The latter problem is considered in [8] and [9] in the case of Example vii, while [8] considers also the problem of minimaxing one component of risk subject to inequalities on *two* others, etc. The discussion of [8], [21], and [12] shows at once on application of our theorem that results of all these types hold for appropriate fixed sample-size procedures, or probability mixtures thereof, in Examples vii to xiv.

5. Sequential problems with continuous time. In this section we will use the method developed in Secs. 3 and 4 to obtain certain sequential minimax results for decision problems concerned with stochastic processes with continuous time parameter. Two types of problems will be considered: in Part I of this section we treat problems where the invariance is present in the same form as in Sec. 4, while in Part II the invariance has to do with the time parameter.

I. *Extension of Section 4 to continuous time.* The problems we consider here will be continuous time analogues of certain of the problems of Sec. 4 (in fact, those of Sec. 4 can be considered as special cases of those here, in the manner of [12]). Since the proofs are essentially identical to those of Sec. 4, we shall not give them. In fact, rather than to state a general theorem, we shall merely list three examples. In each of these the separable process $\{X(t), t \geqq 0\}$ is one of independent and stationary increments which can be taken to be continuous on the right, and $X(T)$ is sufficient for $\{X(t), 0 \leqq t \leqq T\}$. As in Sec. 4, W can be a function of $\theta^{-1} d$ (θ being the unknown parameter) and of the experimentation decision, but for convenience of exposition we discuss the case where it is a sum $W_1 + W_2$. The cost of experimentation W_2 may either be taken to be of the form $W_2(T)$ if the process is observed continuously up to time T, or else the cost may be allowed to depend on the number and spacing of the instants at which the process is observed; a description of this and other modifications

(such as the problem of having to give an estimate continuously), as well as a more detailed discussion of the nature of sequential decision functions in the case of continuous time, and of the processes considered, will be found in [12]. In all of the examples, assumptions on W_2 can be treated as in Sec. 4. The analogue here of the restriction to $\mathfrak{D}^1$ in Sec. 4 is that we must restrict ourselves to the union over all $\epsilon > 0$ of the classes $\mathfrak{D}^\epsilon$ of procedures which observe the process for at least $0 \leqq t \leqq \epsilon$ w.p.1 for all $F \in \mathfrak{F}$. When we consider $\mathfrak{D}^\epsilon$, the g_x is a function of $X(\epsilon)$. As in Sec. 4, it will be easy in most practical cases to decide whether there will be a minimax procedure in $\mathfrak{D}^\epsilon$ for some $\epsilon > 0$ or a minimax procedure which does not observe the process at all.

In each of the three examples, our result is, under assumptions on W like those of Sec. 4, that *there exists an invariant minimax or ϵ-minimax procedure which observes the process for a constant length of time* w.p.1 (or a minimax procedure which does not observe the process at all). Formulas for computing the minimax procedure can be given as in Sec. 4 or [12], and Remark 9 of Sec. 4 applies also to these examples.

(xvi) The process is the one-dimensional Wiener process with known variance per unit time and with $EX(t) = \theta_1 t$, the object being to estimate θ_1. Thus, G, $\mathfrak{F}$, D, and the form of W_1 are the same as in Example vii. In particular, in the special case of a symmetric monotone W_2, we obtain the result of Sec. 5 of [12].

(xvi′) For the Wiener process with unknown scale or unknown location and scale, it has been shown in [12] that the scale parameter can be estimated with arbitrarily high accuracy in arbitrarily short time. Hence, the only new practical problems that arise when the scale parameter is unknown do so because W_2 reflects the number of instants at which the process is observed. In this case, as indicated in [12], we obtain problems analogous to Example xi with G the affine group, or to Example x′ (see also the next example below). In either of these problems there will be an invariant minimax procedure which observes the process at a certain set of instants specified in advance of the experiment.

(xvii) The process is the Gamma process; i.e., $X(0) = 0$ and $X(1)$ has density function $f_{\theta_2\gamma}$ of Sec. 4 with $\theta_1 = 0$ and γ known, the object being to estimate the scale parameter θ_2. Here $\mathfrak{F}$, D, G, and W_1 are the same as in Example x of Sec. 4.

(xviii) Consider the J-variate Wiener process $X(t) = (X_1(t), \cdots, X_J(t))$ where the $X_i(t)$ are independent with known scale factors and $EX_i(t) = \theta_i t$, the θ_i being unknown, $-\infty < \theta_i < \infty$. This is the continuous time analogue of Example xiv, and the considerations there and in xiv+ carry over to the present example.

II. *Invariance in time.* We now consider a process $\{X(t), t \geqq 0\}$ with unknown parameter $\theta > 0$ and with the property that, if $\{X(t), t \geqq 0\}$ has probability law labeled θ, then the process $\{X_c(t), t \geqq 0\}$, defined by $X_c(t) = X(ct)$ where $c > 0$, has probability law labeled $c\theta$. The most familiar process of this kind is the Poisson process. Another such process is the gamma process with θ_2 known and γ unknown.

Suppose the weight function (for estimating θ) is a function only of d_1/θ and

T/θ, where d_1 is the terminal decision and T is the length of time experimentation is carried on (modifications of the type mentioned earlier in this section and discussed in [12] are also possible). Then clearly the multiplicative group of positive reals leaves the problem invariant, where we define $g(\{X(t)\}, \theta, (d_1, T)) = (\{X(gt)\}, g\theta, (gd_1, g^{-1}T))$, the group operation being ordinary multiplication. The difference here from previous problems is that G acts on the process by shifting the time argument of a sample function by a scale factor rather than by operating on the values of the sample function, and that G acts nontrivially on the experimental decision. The reason for allowing this last action and the accompanying dependence of W on T/θ rather than on T lies in the form of the result which this setup yields when one applies the invariance theorem and examines the invariant procedures.

The details here are slightly more delicate and lengthy than those in Part I, so we shall be content with sketching the main idea. Consider the Poisson process with right continuous sample functions. $X(\tau)$ is sufficient for $\{X(t), 0 \leqq t \leqq \tau\}$. Suppose we have a nonrandomized stopping function which depends on the sufficient statistic, i.e., a nonnegative functional T of the process with the property that the event $t_1 < T \leqq t_2$ is measurable with respect to the Borel field generated by $\{X(t), t_1 < t \leqq t_2\}$. For such a T to be invariant we must have $T(x) = cT(x_c)$ for all $c > 0$ and all sample functions x, where x_c is the sample function of X_c when x is the sample function of X. It is easy to see that such a stopping function as $T(x) =$ constant is not invariant, while $T_r(x) =$ first time t that $x(t) = r$, where r is a fixed positive integer, is. In the present problem we must restrict $\mathfrak{D}$ to decision functions which observe the process until at least the first time $X(t) = 1$ (that time gives g_x^{-1}). Under fairly general conditions one can verify whether or not a minimax procedure should observe the process at all and that, if it does, a stopping rule of the type T_r is minimax. Of course, an invariant nonrandomized estimator will be of the form constant/T_r. A special case of this result thus shows that the procedure suggested in Sec. 3 of [22] and which was asserted there to be minimax among all procedures using a particular stopping rule T_r (analogous to a fixed sample-size problem) actually has an optimum property among all sequential procedures: e.g., among all procedures which give at least the prescribed accuracy of estimation, this one minimaxes $E_\theta T/\theta$.

REFERENCES

[1] M. P. Peisakoff, "Transformation Parameters," Thesis, Princeton University, 1950.

[2] E. J. G. Pitman, "The estimation of location and scale parameters of a continuous population of any given form," *Biometrika*, Vol. 30 (1939); "Tests of hypotheses concerning location and scale parameters," *Biometrika*, Vol. 31 (1939), pp. 200–215.

[3] A. Wald, "Contributions to the theory of statistical estimation and testing hypotheses," *Ann. Math. Stat.*, Vol. 10 (1939), pp. 299–326.

[4] G. A. Hunt and C. Stein, "Most stringent tests of statistical hypotheses," unpublished.

[5a] E. L. Lehmann, Notes on testing hypotheses, University of California, Berkeley, 1949.

[5b] E. L. LEHMANN, "Some principles of the theory of testing hypotheses," *Ann. Math. Stat.*, Vol. 21 (1950), pp. 1–26.

[6] M. A. GIRSHICK AND L. J. SAVAGE, "Bayes and minimax estimates for quadratic loss functions," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1951, pp. 53–74.

[7] D. BLACKWELL AND M. A. GIRSHICK, *Theory of Games and Statistical Decision Functions*, John Wiley and Sons, New York, 1954.

[8] J. WOLFOWITZ, "Minimax estimates of the mean of a normal distribution with known variance," *Ann. Math. Stat.*, Vol. 21 (1950), pp. 218–230.

[9] C. STEIN AND A. WALD, "Sequential confidence intervals for the mean of a normal distribution with known variance," *Ann. Math. Stat.*, Vol. 18 (1947), pp. 427–433.

[10] J. L. HODGES, JR., AND E. LEHMANN, "Some applications of the Cramér-Rao inequality," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1951, pp. 13–22.

[11] J. KIEFER, "Sequential minimax estimation for the rectangular distribution with unknown range," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 586–593.

[12] A. DVORETZKY, J. KIEFER, AND J. WOLFOWITZ, "Sequential decision problems for processes with continuous time parameter. Problems of estimation," *Ann. Math. Stat.*, Vol. 24, 1953, pp. 403–415.

[13] P. R. HALMOS, *Measure Theory*, D. Van Nostrand Co., Inc., New York, 1950.

[14] C. CHEVALLY, *Theory of Lie Groups*, Princeton University Press, Princeton, 1946.

[15] A. WALD, *Statistical Decision Functions*, John Wiley and Sons, New York, 1950.

[16] L. LECAM, "An extension of Wald's theory of statistical decision functions," *Ann. Math. Stat.*, Vol. 26 (1955), pp. 69–81.

[17] C. STEIN, "On tests of certain hypotheses invariant under the full linear group" (abstract), *Ann. Math. Stat.*, Vol. 26 (1955), p. 769.

[18] R. R. BAHADUR, "Sufficiency and statistical decision functions," *Ann. Math. Stat.*, Vol. 25 (1954), pp. 423–462.

[19] T. W. ANDERSON, "The integral of a symmetric unimodal function over a convex set and some probability inequalities," *Proc. Amer. Math. Soc.*, Vol. 6 (1955), pp. 170–176.

[20] C. STEIN, "Inadmissibility of the usual estimate for the mean of a multivariate normal distribution" (abstract), *Ann. Math. Stat.*, Vol. 26 (1955), p. 157.

[21] C. R. BLYTH, "On minimax statistical decision problems and their admissibility," *Ann. Math. Stat.*, Vol. 22 (1951), pp. 22–42.

[22] M. A. GIRSHICK, H. RUBIN, AND R. SITGREAVES, "Estimates of bounded relative error in particle counting," *Ann. Math. Stat.*, Vol. 26 (1955), pp. 276–285.

ON THE COMPARATIVE ANATOMY OF TRANSFORMATIONS[1]

By John W. Tukey

Princeton University

1. Summary. The attention of statisticians has usually been focussed on single transformations, rather than on families of transformations. With a growing appreciation of the advantages of examining the behavior of data or approximations over whole families of transformations (Moore and Tukey [2], Anscombe and Tukey [1]), there arises a need for rationally planned charts for representing families of transformations.

The contributions which (i) the topology of the family and (ii) a definition of the strength of a transformation can make to charting are studied in general and applied to the charting of the simple family of transformations. This family is defined to include all transformations of the form

$$y \text{ is replaced by } z = (y + c)^p$$

and all their limits. It thus includes $z = \log(y + c)$, $z = e^{my}$ and the special case

$$z = \begin{cases} 0, & y = y_{\min}, \\ 1, & \text{otherwise}, \end{cases}$$

where $y_{\min}$ is the least value of y either (i) present in the data or (ii) possible, as well as all linear transformations of these transformations.

Experience having shown that transformations with $p \leqq 1$ are much more frequently useful than any others, the charts developed, presented, and exemplified here are restricted to the part of the simple family—its central region—for which $p \leqq 1$. Separate charts are presented for two cases which should cover most cases which arise in practice:

(1) Where, as with counted data and small counts, the least reasonable value for $y + c = 0$, and this value is likely to occur;

(2) Where $y + c$ is always safely > 0, and the range of y is through not many powers of 10.

2. Introduction. In the statistics of today, transformations seem to have two sorts of uses:

(1) providing approximations for theoretical purposes or general convenience;

(2) aiding in the analysis of data by bending the data nearer the Procrustean bed of the assumptions underlying conventional analyses.

Both are interesting and important. In both the quality of the work of the transformation is often judged by the numerical value of some suitable criterion, and transformations which make the value of this criterion sufficiently near some ideal value are acceptable. (Examples follow below.) If we are to consider any

Received July 28, 1955.

[1] Prepared in connection with research sponsored by the Office of Naval Research.

member of a family of transformations as a potentially useful candidate, we should like to understand how one or more criteria vary over the whole family. If we could write a sufficiently simple expression for each criterion in terms of the parameters of the family, we could perhaps reach this understanding algebraically. But this is almost always impossible. We usually have to evaluate each criterion numerically for each of a number of transformations, and then synthesize these few numerical values into an understanding as best we can. If we can find a suitable chart, we can undoubtedly make much more effective use of these few numerical values. Hence the desire for a satisfactory chart.

As one example, note that intermediate statistical texts often prove that, as the number of degrees of freedom $\rightarrow \infty$, the distribution of χ^2 tends to normality. None, to my knowledge, goes on to remark that the same is true of any fixed rational function of χ^2, and hence, in particular, of the results of applying any of the transforms of the simple family. But, knowing this, we are not surprised to find Fisher using the 1/2th power, or Wilson and Hilferty using the 1/3rd power of χ^2 in seeking for a good normal approximation. Nor does the approximate normality of $\log \chi^2$ surprise us very much. If, for a given number of degrees of freedom, we wish to select one transformation of χ^2 to approximate normality, our choice would be greatly eased by charts showing, for example, the deviations of selected percentage points of the approximations from the values appropriate to normality.

Notice how disorganized and unrelated is our information on such approximations. The approximate normalization by transformation of a Poisson distribution of average value λ can be quite well accomplished by two apparently unrelated transformations: $z = \sqrt{y}$ and $z = \log (y + \lambda)$. No connection between these two successful approximations seems to have been recognized, although our definition of the strength of a transformation now suggests a reasonably close connection.

In analyzing data which does not match the assumptions of the conventional methods of analysis, we have two choices [1]. We may bend the data to fit the assumptions by making a transformation. Or we may develop new methods of analysis with assumptions which fit the data in its "original" form somewhat better. If we can find a satisfactory transformation, it will almost always be easier and simpler to use it rather than to develop new methods of analysis. To judge of its satisfactoriness, we need a criterion. The precise nature of the criterion will depend on the situation—and on what ills we are trying to remove by transformation: nonadditivity of effect, nonconstancy of variance, nonnormality of distribution, or what have you.

If we seek to remove nonadditivity of effect, for example, we may take as our criterion the t-value for removable nonadditivity. An example has already been discussed by Moore and Tukey [2]. With a few additional values, this example is used below to illustrate one of the suggested chart forms. In treating other ills, other criteria would of course be involved.

3. Strength and local structure of transformations. The strength of transformations is investigated in general in Sections 5 to 7 and applied to the simple family in Sections 8 to 10. The definition of strength to which we are led is much that the strength of

$$z = y^p$$

is naturally taken as $1 - p$, so that the sequence of transformations

$$z = y,$$
$$z = \sqrt{y},$$
$$z = \log y,$$
$$z = y^{-1/2},$$
$$z = y^{-1},$$
$$\cdots$$

is an equally spaced sequence of strength $0, 1/2, 1, 3/2, 2, \cdots$.

If we compare an arbitrary transformation with these unmodified power transformations, we are led to define its *power strength* as

$$\frac{\log \left[\frac{dz}{dy}\right]_{y_1} - \log \left[\frac{dz}{dy}\right]_{y_2}}{\log (y_2/y_1)}.$$

For $z = y^p$ this definition yields $1 - p$ independent of y_1 and y_2. In general the result will depend on both y_1 and y_2, and may be thought of as the strength of the unmodified power transformation which resembles the given transformation at $y = y_1$ and $y = y_2$.

The philosophy underlying our exploration and heavy use of the local properties of a family of transformations is discussed in Sections 11 and 12. The topologies and differential structures involved are discussed in Sections 13 and 14. The resulting techniques are applied to the body of the simple family in Sections 15 to 20, and to the corners (near the identity and near $z = \text{sgn}^+ y$) in Sections 21 to 24. (The detailed conclusions are summarized in Sections 20, 24, and 25.)

The results of these analyses are then applied in Sections 25 to 27 to the design of the charts.

4. The charts. As noticed in the Summary, the charts are confined to only part of the simple family, i.e., to $z = (y + c)^p$ with $p \leqq 1$ and $0 \leqq y + c$ and their limits. The restriction to positive $y + c$ is easily understood. It ensures monotone-increasing, well-defined, real-valued functions. The restriction to $p \leqq 1$ has no such mathematical reason.

Its origin is entirely empirical and is basically rooted in the psychology of data-gatherers. Experience seems to show quite uniformly that when transformation helps empirical data, it is almost always for $p \leqq 1$. This is clearly a statement

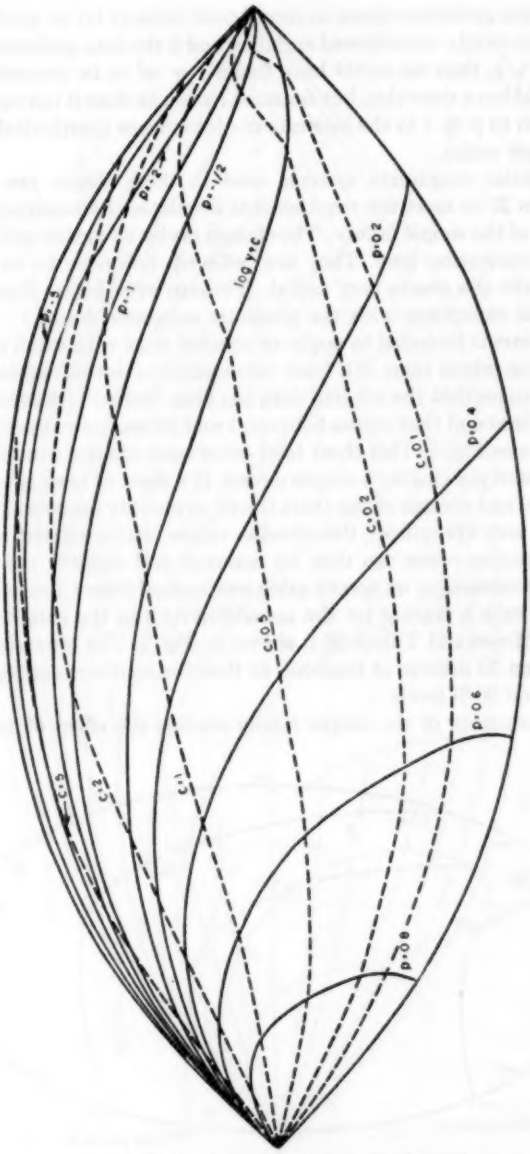

FIG. 1. Basic chart when y is a small count.

about how data-gatherers choose to record (and initially try to analyze) data. For if z is the successfully transformed variable, and if the data-gatherer had chosen to record $w = \sqrt{z}$, then we would have found $z = w^2$ to be successful. The data-gatherer could have done this, but for some reason he does it infrequently. Hence, the restriction to $p \leqq 1$ in the interests of convenience (particularly through the resulting larger scale).

The particular coordinate systems used in these charts are developed in Sections 25 to 27 to meet the requirements developed by studying the topology and strength of the simple family. The choices are by no means unique and do not deserve summarization here. They are believed, however, to be good enough choices to make the charts very useful. (Perhaps even better forms will evolve from practical experience with the presently suggested forms.)

The first chart is intended to apply to counted data with small counts, and to other situations where there is a least value which is not infrequently observed. The chart assumes that the original data has been "coded" (rescaled) so that the least value is zero and that values between 1 and 10 are quite common. The basic chart is shown as Fig. 1. This chart (and other basic charts) can be conveniently used for repeated plotting by a simple device. If a sheet of tracing paper is placed over the chart, and enough of the chart traced to identify locations (for this chart the bounding arcs will suffice), the criterion values can be entered at appropriate points. The tracing paper can then be removed and roughly contoured. (This technique of economizing on special grids was learned from Churchill Eisenhart.) The result of such a tracing for the nonadditivity t in the chinch-bug example discussed by Moore and Tukey [2] is shown as Fig. 2. The numerical entries are values of $10t$ on 20 degrees of freedom, so that the contours correspond roughly to 1%, 5%, and 20% levels.

If a transformation of the simple family renders the effect of treatment and

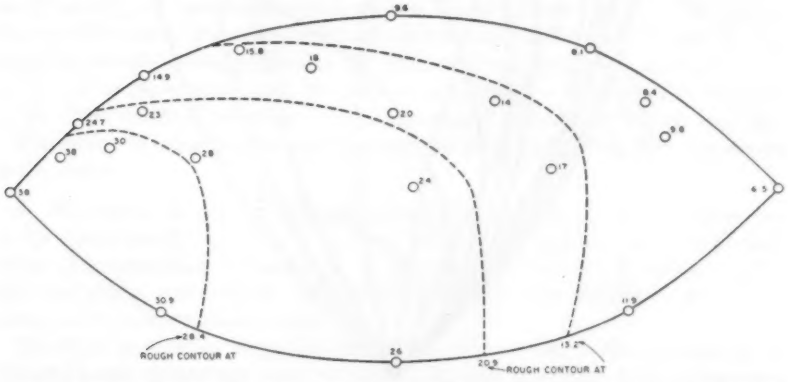

FIG. 2. Use of Fig. 1 to obtain confidence areas. (Entries $10t$, where t measures removable nonadditivity.)

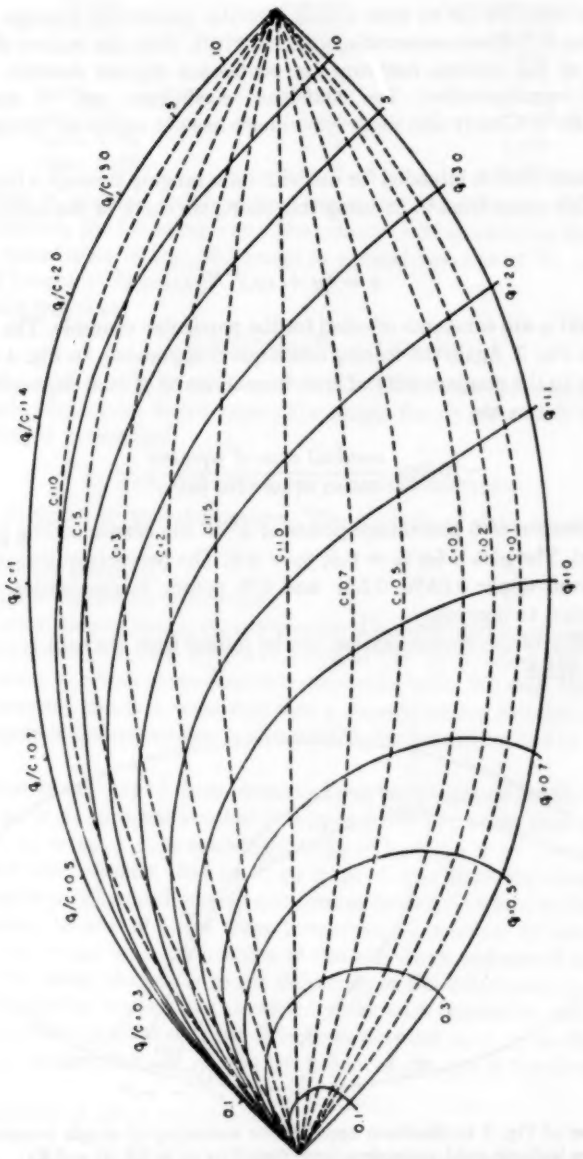

FIG. 3. Basic chart where y is never very large or very small. (Transformation taken as $(y + cy_0^{1-q/q_0})$.)

replication additive (or at least makes zero the population average of the contrast chosen to indicate removable nonadditivity), then the regions above and to the right of the various contours are confidence regions for this additivity-producing transformation. The confidence coefficients are, of course, 99%, 95%, and 80%. Clearly this single experiment has not narrowed things down too much.

The second chart is intended for use with data ranging through a few powers of 10 and safely away from 0. In using this chart, we think of the simple family in the form

$$z = (y + cy_0)^{1-qe_0},$$

where y_0 and e_0 are constants selected for the particular example. The basic chart is shown as Fig. 3. Again the tracing technique is applicable. In Fig. 4 we show an application to the nonnormality of transformations of χ^2 on 6 degrees of freedom. The values shown are

$$-10 \log_{10} \frac{\text{residual sum of squares}}{\text{mean square for slope}} \qquad .1$$

for the regression of 6 percentage points of χ^2 on the corresponding points for a unit normal. The plot is for $e_0 = 0.5$, $y_0 = 0.6$. The percentage points used were the lower and upper 0.05%, 0.5%, and 5% points. Large entries mean close approximation to normality.

The quality of the approximation can be judged from the case $1 - qe_0 = 0.3$, $c = 0$, for which

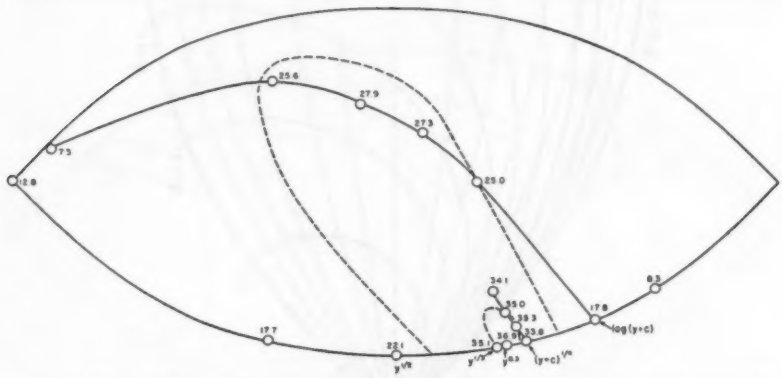

FIG. 4. Use of Fig. 3 to illustrate approximate normality of simple transforms of χ^2_6. (Large entries indicate good approximations. Scaled by $y_0 = 0.6$, $e_0 = 0.5$.)

% point	normality	$3.4300(\chi^2)^{0.30} - 5.6490$
lower 0.05%	−3.291	−3.260
lower 0.05%	−2.576	−2.500
lower 5%	−1.645	−1.674
(50%)	(0.000)	(0.023)
upper 5%	1.645	1.685
upper 0.5%	2.576	2.588
upper 0.05%	3.291	3.261

This quality is perhaps worth noting, since the first attempt to provide an example failed in the following way: The attempt was to examine the normality of simple transforms of w_{10}, the range in normal samples of 10, between the upper and lower 0.01% points.

The result was that

$$4.56446(w_{10})^{.487} - 7.824436$$

agrees with a unit normal to within $\pm 0.01\%$, which is the apparent accuracy of the available cumulative distribution. The simple family often does quite well in transforming to normality!

T. The Strength of Transformations

5. The purposes of transformations. The analysis of data usually proceeds more easily if

(1) effects are additive;

(2) the error variability is constant;

(3) the error distribution is symmetrical and possibly nearly normal.

The conventional purposes of transformation are to increase the degrees of approximation to which these desirable properties hold. We may think of these as three purposes, but it is often true that a transformation suitable for improving one degree of approximation is also suitable for improving one or both of the others.

The failure of any one of these properties can be recognized by the failure of a difference $y_B - y_A$ to equal another difference $y_D - y_C$, as we shall see shortly. Since $y_B - y_A = y_D - y_C$ is equivalent to $y_C - y_A = y_D - y_B$, we may always assume, and shall assume, that $y_A < y_B \leqq y_C < y_D$. Since the change of scale from y to ay is usually (and wisely) accepted as having no effect on the degree of approximation to any of these three properties, it is better to take the ratio $(y_B - y_A)/(y_D - y_C)$ (or some function of this ratio) as a measure of the failure of this property rather than to take the difference of the differences, $(y_B - y_A) - (y_D - y_C)$, since the latter is not invariant under such changes in scale.

We now observe in what way the differences between $y_B - y_A$ on the one hand and $y_D - y_C$ on the other can exhibit the failure of any one of the three desirable properties.

Nonadditivity of effect occurs when

$$y(u_1, v_2) - y(u_1, v_1) \neq y(u_2, v_2) - y(u_2, v_1),$$

where y is a function of the two variables u and v. Usually $y(u, v)$ represents a typical value (population mean, population median, etc.) of the distribution of values obtained for fixed u and v.

Nonconstancy of variability is exhibited when

$$p(u_1) - q(u_1) \neq p(u_2) - q(u_2),$$

where $p(u_1)$ is the p % point of the distribution of observed values when $u = u_1$ and $q(u_1)$ is the corresponding q % point.

Finally, asymmetry of distribution is exhibited when

$$q(u) - y(u) \neq y(u) - p(u),$$

where $p(u)$ and $q(u)$ are upper and lower percentage points cutting off the same tail areas and $y(u)$ is the median, all of the distribution of values observable for a fixed value of u.

In each case, failure is exhibited by a non-zero quartet. From a theoretical point of view our argument is successfully completed. From an empirical one, there remains a gap. Can we use quartets of individual observed values to study observed data as an indication of failures in the population? If we can, then defining strengths in terms of quartets makes very good empirical as well as theoretical sense.

Clearly if we use observed order statistics to estimate percentage points and medians, we can do this for asymmetry and nonconstancy and, if we are willing to deal with population medians, for nonadditivity as well. Since a transform of a sample mean will not be precisely the mean of the transformed values, we cannot expect exact correspondence for nonadditivity defined for population means. If, however, variability for u and v fixed is not large compared with the changes due to alterations in u and v, this disturbance will be rather unimportant and we can regard the means under specified conditions as pseudo-individual values. Since nonadditivity is most important when such additional variability is not too large, it is almost correct, from a practical point of view, to say that the behavior of quartets of individual or pseudo-individual values describes, empirically and adequately, the apparent failures of additivity, constancy of variance, and symmetry of distribution.

We shall, for reasons which will tend to appear as we proceed, take the logarithm of the ratio of differences

$$\log \frac{y_B - y_A}{y_D - y_C} = \log (y_B - y_A) - \log (y_D - y_C)$$

as the measure of the degree of failure of a quartet to correspond to equal differences—whether nonadditivity of effect, nonconstancy of variance, or asymmetry of distribution is involved. There will usually be many such expressions (many quartets $y_A < y_B \leqq y_C < y_D$) which may be worthy of consideration in a particular problem, and transformations will usually be chosen to make these expressions smaller *as a whole*, balancing a gain on one against a loss on another.

6. The strengths of transformations. How shall we assess the strength of a transformation from y to z? Clearly by the extent to which it alters our measures of dissatisfaction. Given a quartet $y_A < y_B \leqq y_C < y_D$, which is transformed into a quartet $z_A < z_B \leqq z_C < z_D$, we have altered

$$\log \frac{y_B - y_A}{y_D - y_C} \quad \text{into} \quad \log \frac{z_B - z_A}{z_D - z_C},$$

and for the initial quartet it is reasonable to measure the strength of the transformation by the change

$$\begin{aligned} k(y_A, y_B; y_C, y_D) &= \log \frac{z_B - z_A}{z_D - z_C} - \log \frac{y_B - y_A}{y_D - y_C} \\ &= \log \frac{(z_B - z_A)(y_D - y_C)}{(z_D - z_C)(y_B - y_A)} \\ &= \log \frac{z_B - z_A}{y_B - y_A} - \log \frac{z_D - z_C}{y_D - y_C}. \end{aligned}$$

The last form of this expression shows that there are natural definitions not only for the results of letting a pair of arguments coalesce

$$k(y_1; y_C, y_D) = k(y_1, y_1; y_C, y_D) = \left[\log \frac{dz}{dy}\right]_{y_1} - \log \frac{z_D - z_C}{y_D - y_C},$$

$$k(y_A, y_B; y_2) = k(y_A, y_B; y_2, y_2) = \log \frac{z_B - z_A}{y_B - y_A} - \left[\log \frac{dz}{dy}\right]_{y_2},$$

but also for the confluent strength arising from the coalescence of both parts

$$k(y_1; y_2) = k(y_1, y_1; y_2, y_2) = \left[\log \frac{dz}{dy}\right]_{y_1} - \left[\log \frac{dz}{dy}\right]_{y_2},$$

all of which involve replacing the difference quotients by their limits, the derivatives of z with respect to y. Clearly we might go one step further, and consider

$$\begin{aligned} k(y) &= \lim_{\Delta \to 0} \frac{k(y; y + \Delta) - k(y, y)}{\Delta} = \lim_{\Delta \to 0} \frac{1}{\Delta} k(y; y + \Delta) \\ &= -\left(\frac{d^2 z}{dy^2}\right) \Big/ \left(\frac{dz}{dy}\right) \end{aligned}$$

as a local measure of the strength of the transformation; however, there are complications which we shall uncover shortly.

With the exception of this last possibility, which we disapprove of, with reasons to be explained, all our measures of the strength of transformations involve at least two arguments (up to as many as four) and the arguments fall into two groups, whose separation (on the y-scale, say) is important. This means that if we face a practical situation, and wish to use a transformation of given strength,

we must begin by selecting one or more quartets typical of the data, and then examine the strength of the proposed transformations for that [those] typical quartets.

7. The composition of transformations. Let us consider successive transformations from y to z and from z to w, together with the composed transformation from y to w. The respective measures of strength are, in an obvious notation:

$$k_{yz}(y_A, y_B; y_C, y_D) = \log \frac{z_B - z_A}{y_B - y_A} - \log \frac{z_D - z_C}{y_D - y_C},$$

$$k_{zw}(z_A, z_B; z_C, z_D) = \log \frac{w_B - w_A}{z_B - z_A} - \log \frac{w_D - w_C}{z_D - z_C},$$

$$k_{yw}(y_A, y_B; y_C, y_D) = \log \frac{w_B - w_A}{y_B - y_A} - \log \frac{w_D - w_C}{y_D - y_C}.$$

It is clear that

$$k_{yw}(y_A, y_B; y_C, y_D) \equiv k_{yz}(y_A, y_B; y_C, y_D) + k_{zw}(z_A, z_B; z_C, z_D)$$

but that, since in general

$$k_{zw}(y_A, y_B; y_C, y_D) \not\equiv k_{zw}(z_A, z_B; z_C, z_D),$$

we will have inequality in

$$k_{yw}(y_A, y_B; y_C, y_D) \not\equiv k_{yz}(y_A, y_B; y_C, y_D) + k_{zw}(y_A, y_B; y_C, y_D).$$

Thus strengths correctly calculated are additive, but strengths naively calculated are not. In particular, if w is the same function of z that z is of y, the strengths of the two transformations (applied successively to the same quartet) will not be the same, and moreover, the strength of the composite will not be twice the strength of either.

Thus we see that the concept of strength is a little more subtle than might be supposed.

In particular, an attempt to use $k(y)$ directly fails because by going to the limit *after division by* Δ (which we must do to obtain a finite limit) we have lost the distinction between the original and the transformed quartet. In more detail, and working in terms of confluent strengths and in terms of $z = f(y)$,

$$k_{yw}(y; y + \Delta) = k_{yz}(y; y + \Delta) + k_{zw}(z; z + \delta),$$

where $z + \delta = f(y + \Delta)$, so that, on division by Δ and passage to the limit,

$$k_{yw}(y) = k_{yz}(y) + \frac{df(y)}{dy} k_{zw}(z)$$

$$= k_{yz}(y) + \frac{dz}{dy} k_{zw}(z),$$

and both the shift in argment and the multiplication by dz/dy must be taken into account.

As an example, consider $z = \sqrt{y}$ and $w = \sqrt{z}$, so that $k_{ys}(y) = 1/2y$ $k_{sw}(z) = 1/2z$, and $dz/dy = 1/2\sqrt{y}$. We obtain, from the formula above,

$$k_{yw} = \frac{1}{2y} + \frac{1}{2\sqrt{y}}\left(\frac{1}{2z}\right) = \frac{1}{2y} + \frac{1}{2\sqrt{y}}\left(\frac{1}{2\sqrt{y}}\right) = \frac{3}{4y},$$

which is exactly what we would have obtained by direct calculation. Note that we combined $1/2y$ with $1/2z$ to obtain $3/4y$, which is far from the naive answer.

II. Strengths in the Simple Family

8. Confluent strengths of powers. We now return to the simple family, and begin with the confluent strengths of the pure powers, where

$$z = \begin{cases} y^p, & p \neq 0, \\ \log y, & p = 0. \end{cases}$$

Here we have

$$k(y_1\,;\,y_2) = \left[\log \frac{dz}{dy}\right]_{y_1} - \left[\log \frac{dz}{dy}\right]_{y_2},$$

where

$$\log \frac{dz}{dy} = \begin{cases} \log p + (p-1)\log y, & p \neq 0, \\ -\log y, & p = 0, \end{cases}$$

so that

$$k(y_1\,;\,y_2) = (p-1)\log(y_1/y_2) \qquad \text{for all } p.$$

Thus the strength is proportional to $p - 1$ for any pair of confluent arguments y_1 and y_2.

If $w = z^q$, and $z = y^p$, we have

$$\begin{aligned} k_{ys}(y_1\,;\,y_2) &= (p-1)\log(y_1/y_2), \\ k_{sw}(z_1\,;\,z_2) &= (q-1)\log(z_1/z_2) = (q-1)\,p\log(y_1/y_2), \\ k_{yw}(y_1\,;\,y_2) &= (pq-1)\log(y_1/y_2) \end{aligned}$$

where we have used

$$\log(z_1/z_2) = p\log(y_1/y_2).$$

We note that the additive decomposition of strengths is, in terms of the numerical coefficients,

$$pq - 1 = (p-1) + p(q-1);$$

so that if $p = \frac{1}{2}$, for example, we again get

$$-\tfrac{3}{4} = -\tfrac{1}{2} + (-\tfrac{1}{4}),$$

showing how the second square-rooting had only half the strength of the first.

The special case $q = 0$ (logarithm for the second transformation) offers no difficulty, the decomposition being

$$-1 = (p - 1) + (-p);$$

but an attempt to put $p = 0$ (logarithm for the first transformation) fails, since in this case

$$\log (z_1/z_2) \neq p \log (y_1/y_2)$$

and the power strengths are quite dependent on the particular values of y_1 and y_2 involved. The resulting transformations, such as $\sqrt{\log y}$, are not power transformations and seem to have been relatively little used.

In particular, the sequence of transforms

$$\cdots, \quad y, \qquad \sqrt{y}, \qquad \log y, \qquad 1/\sqrt{y}, \qquad 1/y, \cdots$$

are seen to be equally spaced in strength, where a more naive approach might have led to $y, y^{1/2}, y^{1/4}, y^{1/8}, \cdots$ as an equally spaced sequence—a possibility not in accordance with either a careful analysis or empirical experience.

In general, we shall refer to the ratio of $k_{y^p}(y_1, y_2)$ to $-\log (y_1/y_2)$ as the *power strength* of a transformation. With this definition, the power strength of y^p is $1 - p$ and the power strengths of the sequence $\sqrt{y}$, $\log y$, $1/\sqrt{y}, \cdots$ are $1/2$, 1, $3/2$, 2, $5/2, \cdots$. (For a transformation which is not a pure power, the power strength will depend on the exact arguments y_1 and y_2.)

9. Confluent strengths in the simple family. We can now go on to consider the confluent strengths of the modified power transformations

$$z = \begin{cases} a_p(y + c)^p, & p \neq 0, \\ \log (y + c), & p = 0, \end{cases}$$

where we are likely to choose the constant a_p for each p so that $\operatorname{sgn} a_p = \operatorname{sgn} p$. We have

$$\log \frac{dz}{dy} = (p - 1) \log (y + c) + \text{constant}$$

and hence

$$k(y_1; y_2) = (p - 1) \log \frac{y_1 + c}{y_2 + c}.$$

The power strength will be given by

$$\frac{k(y_1; y_2)}{-\log (y_1/y_2)} = (1 - p) \frac{\log ((y_1 + c)/(y_2 + c))}{\log (y_1/y_2)}.$$

If we put $y_1 = y - \delta$, $y_2 = y + \delta$, and expand all logarithms in terms of δ, we obtain

$$(1 - p)\left[\frac{y}{y + c} - \frac{\delta^2}{3}\frac{c(2y + c)}{y(y + c)^3} + \cdots\right],$$

where the first term will usually be controlling, and where, by Rolle's theorem, we may slightly redefine y, keeping it well between y_1 and y_2 so that the first term is exact.

We are now able to give some explanation of the fact, mentioned earlier, that $\sqrt{y}$ and $\log(y + \lambda)$ are both fairly good normalizing transformations for a Poisson variate of average value λ. Their respective power strengths at $y = \lambda$ are exactly

$$(1 - \tfrac{1}{2}) = \tfrac{1}{2}$$

and approximately

$$(1 - 0)\frac{\lambda}{\lambda + \lambda} = \tfrac{1}{2},$$

so that their rough equivalence is assured.

A further appeal to Rolle's theorem shows that nonconfluent strengths are always equal to confluent strengths for points near the center of each pair, and it is easy to convince ourselves that, if ratios of not much more than 10 to 1 are involved for y_B/y_A and y_D/y_C, the confluent strengths give us an adequate picture of the behavior of the modified power transformations. This disposes of most practical situations except one. When y is a count, and zero values actually occur, we obtain infinite ratios for y_B/y_A. A brief inquiry into strengths in this cases does not provide any results of much practical use, which is somewhat unfortunate.

10. Extreme strengths. We have expressed our strengths on a logarithmic basis; this is quite reasonable for weak and moderate transformations, but can easily be misleading for extremely strong transformations in practical applications. For example, suppose that only three different values are ever observed, 0, 1, and 2, and that they have been transformed into 0.0000, 1.0000, and 1.0001.

By our definitions, the transformation carrying them into 0.0000 0000, 1.0000 0000, and 1.0000 0001 would be twice as strong. From most practical points of view this is quite wrong, since both transformations are so very close to the one carrying 0, 1, 2 into 0, 1, 1 (which by our definitions has infinite strength). In practice, infinite strength is *not* infinitely far away.

This means that, in charting, we should pay very considerable attention to the behavior of strengths on our scale where they are small or moderate, much less attention to their behavior where they are large, and should avoid moving infinite strengths to infinity.

III. Local Structures of Familie of Transformations

11. Introduction. It is natural to question the intrusion of topology and differential structure into the present paper. For in choosing transformations for empirical purposes—whether for data analysis or theoretical manipulation—only a limited degree of detail about a transformation is relevant. If our interest is coarse-grained, while the attention of topology is directed to the indefinitely fine grain of arbitrarily small neighborhoods, how can topology help us? There are at least two answers to this question, and both are quite relevant.

First, there is a heuristic principle that the study of behavior in the small often, but not always, leads to results which are helpful at a larger scale. In the present circumstances we may reasonably expect this principle to apply without exception.

Second, we are trying to chart the family of transformations in such a way that, when we plot values of some interesting quantity on the chart, we can use these values as advantageously as possible in:

(1) understanding the relation of the transformation to the quantity;

(2) picking a single useful transformation; or

(3) indicating a region of acceptable transformations.

For any and all of these ends, we should like to have the quantities of interest vary smoothly over the chart.

The finest-grained aspect of smooth variation is continuity. Thus we should at least like to have our quantities continuous functions over the chart. To do this we must respect the topology of the family of transformations in choosing the topology of the chart. This respect need only be one-sided, however, for if

(1) position in the family is a continuous function of position in the chart, and

(2) the quantity is a continuous function of position in the family,

then the quantity will be a continuous function of position on the chart. In a phrase, we must not bring together in the chart things which are apart in the family, although we may, if we wish, keep apart in the chart things which come together in the family.

For all that the second answer leaves us free to pull things apart in the chart wherever and as much as we wish, the first answer suggests that we should restrain such tendencies as much as may be reasonable.

12. Intersection and tangency. Whether one curve intersects (we shall always mean "intersects at a non-zero angle" when we say "intersects"), or is instead tangent to, another curve in the family is a relatively easy question to settle, and one that can usefully be asked near almost any point of the family. The analog of the rule just stated above may be put as follows: If the curves intersect in the family, we cannot allow them to be tangent on the chart; if they are tangent in the family, then we can, if we wish, let them intersect on the chart.

Such considerations may arise in the interior of a family, but are more likely to be important on its boundary. If a family is two-dimensional (as the special family is), the boundary will consist of curves and corners. It is at these corners,

which are often singularities in some other sense, that intersection and tangency will be most critical.

At such a corner we should avoid, at all costs, the introduction of new tangencies, and should try to preserve the tangencies of the family whenever this can be done without contravening more important considerations.

We shall have to deal with a least one corner where there are several alternate differential structures. This is likely to force us to eliminate tangencies. For if we have two curves which are tangent in all of these structures, they may not be tangent in the same way ("corresponding" pairs of points, one on each curve, in one structure may not "correspond" in another). In this case, no one sort of tangency in the chart can possibly be satisfactory. We must destroy the tangency and chart the curves as intersecting at a finite angle.

13. Nature of the topology. The discussion above has assumed a topology without describing it. We need to say something more. If we have a family of transformations depending on parameters $\alpha, \beta, \gamma, \cdots$,

$$z = z(y; \alpha, \beta, \gamma, \cdots),$$

and if $\alpha = \alpha(t), \beta = \beta(t), \gamma = \gamma(t), \cdots$ defines a curve of transformations

$$z_t = z(y; \alpha(t), \beta(t), \gamma(t), \cdots),$$

when do we say that z_t converges to z_0 as $t \to 0$? With a mild reservation, we say that $z_t \to z_0$ if there exist A_t and B_t so that

$$A_t + B_t z_t(y) \to z_0(y)$$

for every relevant y. Notice two things here: First, we have to introduce the variable linear transformation defined by (A_t, B_t), since we regard the class of all $C + Dz_t$, for C and D fixed, as equivalent and our topology really refers to equivalence classes of transformations. Second, the convergence is for all relevant y, so that a change in the set of y's considered relevant may change the topology. In our present case—the simple family—such changes only occur (i) near the far corner, or (ii) for degenerate cases where at most two values of y are relevant, and all non-degenerate transformations are equivalent. (A transformation is degenerate if all its values for relevant y's are the same. The mild restriction mentioned above is that z_0 should not be degenerate.)

14. Nature of the differential structure. We have also talked of "intersection" and "tangency" without specific basis. We have available a differential structure in which these terms are well defined, but it is rather different from those structures which appear in, say, Riemannian geometry. It does not start with something like a differential form which gives local meaning to ratios of distances and from which local differentiation, tangency, etc., follow. It cannot, since we shall not have local distances. Instead, it starts with the notion of a direction from one transformation to another and proceeds to the definition of a tangent.

Given two transformations, z_1 and z_2, their difference, $z_1 - z_2$, is also a trans-

formation and is, of course, equivalent to any multiple of itself. If we replace z_1 and z_2 by equivalent transformations, say $A + Bz_1$ and $C + Dz_2$, the difference becomes

$$A + Bz_1 - C - Dz_2 = A - C + B(z_1 - z_2) + (B - D)z_2$$
$$= A - C + (B - D)z_1 + D(z_1 - z_2),$$

namely, any member of a one-parameter family of equivalence classes where $z_1 - z_2$ is additively combined with various amounts of z_1 or z_2. The resulting direction is defined, but not too precisely.

If, however, $z_1 = z_t$ and $z_2 = z_0$ where z_t converges to z_0 as $t \to 0$, then we can focus our attention on a particular $A_t + B_t z_t$ whose values converge to those of z_0. If the convergence of the values is sufficiently differentiable for each y, then we have

$$A_t + B_t z_t(y) = z_0 + tz'(y) + O(t^2),$$

where $O(t^2)$ means terms of order t^2 or smaller as $t \to 0$. We may regard z' considered as a transformation, as the derivative of z_t at the point (at the transformation) z_0.

It is well to note that z' is also not uniquely defined. For if we replace B_t by $B_t + tC$, as we may without disturbing the convergence, then

$$A_t + (B_t + tC)z_t(y) = z_0 + t[z'(y) + Cz_0(y)] + O(t^2),$$

and we see that $z' + Cz_0$ is also a "derivative." Except for the inevitable linear transformations this is the most general form, and we may say that the directions at z_0 correspond to the families of transformations of the form

$$A + Bz' + Cz_0,$$

where A, B, and C are arbitrary constants with $B \neq O$.

Two curves

$$x_t = z_0 + tz' + O(t^2)$$

and

$$z_s = z_0 + s\dot{z} + O(s^2)$$

will have the same direction at z_0 if

$$A + B\dot{z} + Cz_0 \equiv z'$$

for the relevant values of y and suitable constants A, B, and C.

Notice the appearance again of the relevant values of y. It will again usually be true that, so long as y takes at least three different values, it will not matter which three. But in exceptional circumstances it will matter, and it is in this way that a corner may receive various differential structures.

The same sort of construction can be extended to higher derivatives. We shall need it in only one special case, and no general discussion seems necessary.

With these general principles and techniques in mind, then, we can proceed to study the simple family.

IV. The Body of the Simple Family

15. Definitions and limiting forms. As previously stated, the *simple family* of transformations consists of all transformations, which can be obtained by compounding linear transformations with one (integral or fractional) power transformation, and of all limiting forms of such transformations.

The effect of an additional linear transformation from z to $A + Bz$ with $B \neq 0$ applied *last* is always regarded as trivial. Transformations which can be trivially converted into each other are equivalent. Thus $z = 3y^2 - 17$ and $z = 0.001 + 0.037y^2$ are regarded as equivalent to each other and to $z = y^2$. This is not the case for an *initial* linear transformation. Thus $z = \log y$ and $x = \log (y + 10\,000\,000)$ are quite different transformations. When we are dealing with a curve of transformations, such as, for example, the curve of pure power transformations with powers between 1 and 0, namely

$$\{y^p, 1 > p > 0\},$$

we shall not distinguish equality from equivalence, and shall freely write such "equations" as

$$\{3y^p, 1 > p > 0\} = \{y^p, 1 > p > 0\} = \{14 - 2y^p, 1 > p > 0\}.$$

We are only concerned with real values, and prefer monotone functions, so that we could take as our definition of power

$$y^p = \operatorname{sgn}(y) \mid y \mid^p,$$

where the signum function, sgn (y), is defined by

$$\operatorname{sgn}(y) = \begin{cases} +1, & y > 0, \\ 0, & y = 0, \\ -1, & y < 0. \end{cases}$$

With this definition,

$$(y^p)^q = y^{pq}$$

but

$$\frac{dy^p}{dy} = py^{p-1}(\operatorname{sgn} y) = py^{p-1}(\operatorname{sgn} y^p) = py^{p-1}(\operatorname{sgn} y^{p-1} = p \mid y \mid^{p-1}.$$

(Differences from other definitions of powers only occur for negative y, and since negative y's occur infrequently in practice, this possible definition should be regarded as precautionary rather than important. We shall not use it explicitly here, merely remarking that its use would make only trivial alterations in our results.)

The limiting forms are associated with the exceptional values of the power,

$-\infty$, 0, and $+\infty$, and with the extreme values, $-\infty$ and $+\infty$, of the constant. For $p \to 0$, we have

$$(y + c)^p = e^{p \log (y+c)} \approx 1 + p \log (y + c),$$

which, after a suitably varying linear transformation, converges to $\log (y + c)$. For $p \to \pm\infty$ and c tending to a finite limit, the convergence is to less familiar objects. The transformations involved depend on the particular set of values transformed through the extreme values, $y_{\max}$ and $y_{\min}$, actually present in the data. If we define a function φ_0 by

$$\varphi_0(u) = \begin{cases} 0, & u \neq 0, \\ 1, & u = 0, \end{cases}$$

we then have for $p \to +\infty$ and a suitable choice of A_p and B_p,

$$A_p + B_p(y + c)^p \to -\varphi_0(y - y_{\min}) = -1 + \operatorname{sgn}^+(y - y_{\min}),$$

where the last equality is valid for actually occurring y's, since $y < y_{\min}$ is impossible.

As in the sequel, a "+" on a function indicates that negative values are to be replaced by zero. Hence, in particular,

$$\operatorname{sgn}^+ y = \begin{cases} 0, & y \leqq 0, \\ 1, & y > 0. \end{cases}$$

There remains the case where $p \to \pm\infty$ but p/c tends to a finite limit. (Hence $c \to \pm\infty$, also.) Here

$$(y + c)^p = c^p \left(1 + \frac{y}{c}\right) p = c^p \left[\left(1 + \frac{y}{c}\right)^c\right]^{p/c};$$

and since the expression in [ ] tends to e^y, the whole expression, after a suitably varying linear transformation, tends to

$$e^{my},$$

where $m = \lim p/c$.

Thus the simple family contains:

$$z = (y + c)^p,$$

$$z = \log (y + c),$$

$$z = e^{my},$$

and all linear transforms of these transformations.

16. Normalization. Because (i) they are apparently of the greatest practical importance, and (ii) they determine the remaining cases by symmetry, we shall confine our detailed analysis to transformations with $p \leqq 1$ (and their limits). Consequently, we shall be concerned with least values but not with greatest

ones. It will therefore be convenient, for the remainder of this part, for us to fix initial scale of the y's so that the three smallest values of y are 0, 1, and y_2, where $y_2 > 1$. This can be done by a linear transformation, and undone again by another, so that our conclusions are to be changed by at most a linear transformation when we are through.

We shall also assume that the reasonable values of c are such that $y + c$ is non-negative. We shall treat the general case, specializing later to the case where c is positive and bounded away from zero.

17. Tangency and intersection for c small. We now investigate the situation along and near the curve $\{z = y^p\}$ in more detail. We begin by considering $(y + c)^p$ as $c \to 0$, where the value of p is important. If $p < 0$, and temporarily we shall write $p = -k$, then

$$(y + c)^p = \begin{cases} c^p = c^{-k} \to \infty, & y = 0, \\ (y + c)^{-k} \to y^{-k}, & y > 0, \end{cases}$$

and

$$1 - c^k(y + c)^p = \begin{cases} 0, & y = 0, \\ 1 - \left(\dfrac{c}{y + c}\right)^k, & y > 0 \end{cases}$$

$$= \operatorname{sgn}^+(y) - c^k(y^{-k} \operatorname{sgn}^+(y)) + O(c^{k+1}).$$

As $c \to 0$ for k fixed, the transformation tends to $\operatorname{sgn}^+(y)$ and appears like an additive mixture of this and $y^{-k}\operatorname{sgn}^+(y)$. This additive piece is different for different values of k, so that we learn that the curves of transformations

$$\{(y + c)^p, c \to 0\} = \{1 - c^k(y + c)^p, c \to 0\}, \qquad p < 0,$$

tend to $\operatorname{sgn}^+(y)$, but no two have a common tangent.

For $p = 0$, we must deal with $\log (y + c)$, where

$$1 - \frac{\log (y + c)}{-\log c} = \begin{cases} 0, & y = 0, \\ 1 - \dfrac{\log^+(y + c)}{-\log c}, & y \geqq 1, \end{cases}$$

$$= \operatorname{sgn}^+ y - \frac{\log^+(y + c)}{-\log c}$$

$$= \operatorname{sgn}^+ y - \frac{\log^+ y}{-\log c} + O\left(\frac{c}{-\log c}\right),$$

and hence the curve

$$\{\log (y + c), c \to 0\} = \left\{1 - \frac{\log (y + c)}{-\log c}, c \to 0\right\}$$

also tends to $\operatorname{sgn}^+ y$, but with a different tangent.

There is also another simply described curve with c small (in fact with $c = 0$) which tends to $\operatorname{sgn}^+ y$. This is the curve

$$\{y^p,\ p \to 0\},$$

where we have

$$y^p = \begin{cases} 0, & y = 0, \\ y^p = e^{p \log y}, & y > 0, \end{cases}$$
$$= \operatorname{sgn}^+ y + p \log^+ y + O(p^2), \qquad y \geqq 0.$$

We see now that the curves

$$\{y^p,\ p \to 0\} \text{ and } \{\log(y + c),\ c \to 0\}$$

are tangent to each other at the transformation ($z = \operatorname{sgn}^+ y$), with p corresponding to $1/(-\log c)$.

For $p > 0$, we have

$$(y + c)^p = \begin{cases} c^p, & y = 0 \\ y^p \left(1 + \dfrac{c}{y}\right)^p, & y > 0, \end{cases}$$
$$= \varphi_0(y) c^p + y^p + cpy^{p-1} + O(c^2),$$

so that the curve

$$\{(y + c)^p,\ c \to 0\}$$

tends to the transformation y^p.

Near y^p we are also interested in

$$y^{p+\delta} = y^p e^{\delta \log y} = y^p + \delta y^p (\log y) + O(\delta^2),$$

which is easily seen to have a different tangent than the previous curves at y^p.

18. Tangency and intersection for large c. If c is large,

$$(y + c)^p = c^p \left(1 + \frac{y}{c}\right)^p = c^p \exp\left[p \log\left(1 + y/c\right)\right]$$
$$= \exp\left[(py/c) + py^2/2c^2 + O(1/c^3)\right]$$
$$= c^p e^{py/c} \left(1 + \frac{py^2}{2c^2} + O\left(\frac{1}{c^3}\right)\right).$$

Thus, if $c = 1/\epsilon$ and $p = m/\epsilon$ (note that $m < 0$ if p is always < 0), we have

$$c^{-p}(y + c)^p = e^{my} + \frac{\epsilon}{2} my^2 e^{my} + O(\epsilon^2 e^{my}),$$

and the curve

$$\{(y+c)^p,\ p = mc,\ c \to \infty\} = \{c^{-p}(y+c)^p,\ p = mc,\ c \to \infty\}$$

tends to e^{my}.

We also wish to consider $\{e^{(m+\delta)y},\ \delta \to 0\}$ for which

$$e^{(m+\delta)y} = e^{my}e^{\delta y} = e^{my}(1 + \delta y + O(\delta^2))$$
$$= e^{my} + \delta y e^{my} + O(\delta^2 e^{my}).$$

Thus the curves

$$\{(y+c)^p,\ p = mc,\ c \to \infty\} \qquad \text{and} \qquad \{e^{(m+\delta)y},\ \delta \to 0\}$$

tend to the same limit, but with different tangents.

We also have to consider $c \to \infty$ and p fixed, for which

$$(y+c)^p = c^p\left(1 + \frac{y}{c}\right)^p = c^p\left(1 + \frac{py}{c} + \frac{p(p-1)}{2c^2}\,y^2 + O\left(\frac{y}{c}\right)^3\right)$$

and

$$\frac{c}{p}\,c^{-p}(y+c)^p - \frac{c}{p} = y + \frac{p-1}{2c}\,y^2 + O\left(\frac{1}{c^2}\right),$$

so that the curves

$$\{(y+c)^p,\ c \to \infty\} = \left\{\frac{c}{p}\,c^{-p}(y+c)^p - \frac{c}{p},\ c \to \infty\right\}$$

all tend to y with a common tangent, a natural parameter along this tangent being $(p-1)/2c$.

Since e^{my} has been included, we must also consider the case $m \to 0$ where

$$\frac{e^{my}-1}{m} = y + \frac{m}{2}\,y^2 + O(m^2),$$

so that

$$\{e^{my},\ m \to 0\}$$

has the same tangent as the curves just considered, with m corresponding to $(p-1)/c$.

19. Tangency and intersection for c moderate. If $p \to 1$ with c constant, we learn from

$$(y+c)^{(1+\delta)} = (y+c)e^{\delta\log(y+c)} = (y+c)(1 + \delta\log(y+c) + O(\delta^2))$$

that

$$(y+c)^{(1+\delta)} - c = y + \delta(y + c\log(y+c) + O(\delta^2)$$

and hence that the curves

$$\{(y + c)^p,\ p \to 1\}$$

tend to y, each with its own tangent.

At the other extreme, where $p = -k$ tends to $-\infty$, we have

$$1 - c^k(y + c)^p = \begin{cases} 0, & y = 0, \\ 1 - \left(1 + \frac{1}{c}\right)^{-k}, & y = 1, \\ 1 - \left(1 + \frac{1}{c}\right)^{-k}\left(\frac{c+1}{c+y}\right)^k, & y > 1, \end{cases}$$

$$= \operatorname{sgn}^+ y - \left(1 + \frac{1}{c}\right)^{-k} \varphi_0(y - 1) + O\left(\left(\frac{c+1}{c+y}\right)^k\right);$$

hence the curves

$$\{(y + c)^p,\ p \to -\infty\} = \{1 - c^k(y + c)^p,\ p \to -\infty\}$$

tend to $\operatorname{sgn}^+ y$ with a common tangent and a natural parameter which is a multiple of $[1 + (1/c)]^{-k} = [1 + (1/c)]^p$.

We also have, as $m \to -\infty$,

$$1 - e^{my} = \begin{cases} 0, & y = 0, \\ 1 - e^m, & y = 1, \\ 1 - e^m e^{m(y-1)}, & y > 1, \end{cases}$$

$$= \operatorname{sgn}^+(y) - e^m(\varphi_0(y - 1)) + O(e^{m(y-1)}),$$

and we see that the curve

$$\{e^{my},\ m \to -\infty\} = \{1 - e^{my},\ m \to -\infty\}$$

has the same limit and tangent, with e^m playing the role of $[1 + (1/c)^p$.

Everything is well behaved as $p \to 0$ when $(y + c)^p$ tends to $\log (y + c)$.

20. The resulting picture. If we put together all the individual results of the last three sections, we are forced to the qualitative relation of the curves

$\{e^{my}, m < 0\}$,

$\{y,^p 1 > p > 0\}$,

$\{(y + c)^p, p \text{ fixed } (1 > p > 0), 0 < c < \infty\}$,

$\{\log (y + c), 0 < c < \infty\}$,

$\{(y + c), p \text{ fixed } (p >), 0 < c < \infty\}$,

$\{(y + c)^p, c \text{ fixed}, 1 > p > -\infty\}$

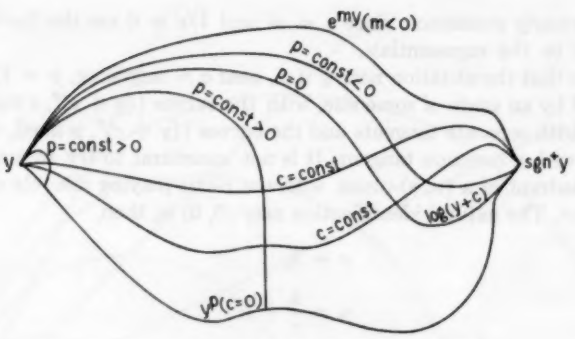

FIG. 5. The topology of the simple family.

shown in Fig. 5. These relations of limit, tangency, and intersection substantially limit the diagrams on which we may reasonably plot the transformations $(y + c)^p$ (with their limits) for $1 \geqq p > -\infty$ and $0 \leqq c < \infty$.

V. THE CORNERS OF THE SIMPLE FAMILY

21. The neighborhood of the near corner. We can try to learn more about the situation near the corners at y and $\text{sgn}^+ y$ by carrying the expansion out to further terms. We begin with the neighborhood of y, where

$$\frac{c}{p}\, c^{-p}(y + c)^p = y + \frac{p - 1}{2c}\, y^2 + \frac{(p - 1)(p - 2)}{6c^2}\, y^3 + O\left(\frac{1}{c^3}\right),$$

$$\frac{e^{my} - 1}{m} = y + \frac{m}{2}\, y^2 + \frac{m^2}{6}\, y^3 + O(m^3),$$

$$(y + c)^{1+\delta} - c = y + \delta(y + c) \log (y + c) + \frac{\delta^2}{2}\, (y + c)(\log (y + c))^2 + O(\delta^3),$$

and where, in particular, we may put $c = 0$ in the last expansion.

If, taking account of the signs of $p - 1$ and m in the region we are interested in, we set

$$h = \frac{1 - p}{c} = -m,$$

the first two expansions give

$$\frac{c}{p}\, c^{-p}(y + c)^p = y + h\left(-\frac{y^2}{2}\right) + h\left(\frac{1}{c} + h\right)\left(\frac{y^3}{6}\right) + O\left(\frac{1}{c^3}\right),$$

$$\frac{1}{m}\, (e^{my} - 1) = y + h\left(-\frac{y^2}{2}\right) + h(O + h)\left(\frac{y^3}{6}\right) + O\left(\frac{1}{c^3}\right),$$

which are clearly consistent, since $c = \infty$ and $1/c = 0$ are the limiting values which lead to the exponentials.

We know that the situation near y (i.e., near $c =$ anything, $p = 1$) should be represented by an angle of some size, with the curves $\{(y + c)^p, c \text{ fixed}, p \to 1\}$ coming in with separate tangents and the curves $\{(y + c)^p, p \text{ fixed}, c \to \infty\}$ all coming in with a common tangent. It is not unnatural to try to map this into the first quadrant of a (u, v)-plane, with the v-axis playing the role of the common tangent. The natural identification near (0, 0) is, then,

$$v = h,$$

$$u = \frac{h}{c} + h^2.$$

However, experimentation leads to complexities, and reflection shows us that we may reasonably alter the h^2 term, since it is a higher-order function of the h term.

Omitting the h^2 term,

$$u = \frac{h}{c}, \qquad v = h$$

whence

$$c = \frac{v}{u},$$

and the radius of curvature at (0, 0) of the curves $p =$ constant may be found from the parametric form for a circle through (0, 0) tangent to the v-axis, namely,

$$v = R \sin \psi \sim R\psi,$$

$$u = R(1 - \cos \psi) \sim \tfrac{1}{2}R\psi^2,$$

whence

$$R = \frac{(R\psi)^2}{2(R\psi^2/2)} \sim \frac{V^2}{2u} = \frac{h^2}{2h/c} = \frac{hc}{2} = \frac{1 - p}{2}.$$

We thus reproduce the situation around y to this accuracy if we make the following identifications

$$\{\text{curve of constant } c\} = \{\text{ray through origin of slope } c\},$$

$$\{\text{curve of constant } p\} = \{\text{circle with (0, 0) and } (1 - p, 0) \text{ as diameter}\}.$$

We then have a map of the simple family which is sound for transformations near the identity.

It is this map which was used by Moore and Tukey [2] and is shown as Fig. 6. In polar coordinates, $\tan \theta = c$ and $r = (1 - p)/\sqrt{1 + c^2}$.

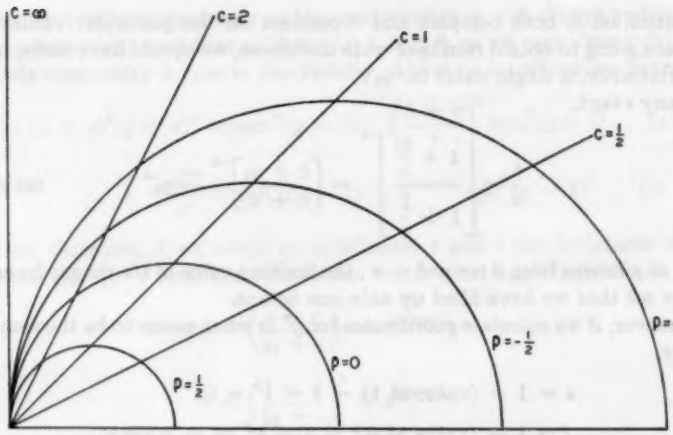

FIG. 6. The neighborhood of the near corner.

22. The neighborhood of the far corner. Now we go to the neighborhood of $\operatorname{sgn}^+ y$, and again carry more terms to find

$$1 - c^k(y + c)^p = \operatorname{sgn}^+ y - \left(1 + \frac{1}{c}\right)^{-k} \varphi_0(y - 1)$$
$$- \left(1 + \frac{y_2}{c}\right)^{-k} \varphi_0(y - y_2) + \cdots \qquad (p = -k \to -\infty),$$

$$1 - e^{my} = \operatorname{sgn}^+ y - e^m \varphi_0(y - 1) - e^{my_2} \varphi_0(y - y_2) + \cdots \qquad (m \to \infty),$$

$$1 - c^k(y + c)^p = \operatorname{sgn}^+ y - \left(\frac{y}{c}\right)^{-k} \operatorname{sgn}^+ y + \cdots \qquad (p = -k < 0, c \to 0),$$

$$1 - \frac{\log (y + c)}{-\log c} = \operatorname{sgn}^+ y - \frac{\log^+ y}{\log c} + \cdots \qquad (c \to 0),$$

$$y^p = \operatorname{sgn}^+ y + p \log^+ y + \cdots \qquad (p \to 0).$$

We again concentrate on the curves which come to a common tangent. (At this corner these are the family c = constant rather than the family p = constant.) This directs our attention to the first of the five formulas and to the quantities

$$s = \left(1 + \frac{1}{c}\right)^{-k} \quad \text{and} \quad t = \left(1 + \frac{y_2}{c}\right)^{-k},$$

so that we would have

$$\frac{1}{c} = s^k - 1$$
$$y_2 = \frac{t^k - 1}{s^k - 1}.$$

The situation is both complex and dependent on the particular values of y_2. If we are going to obtain results of wide usefulness, we should have to fix, more or less arbitrarily, a single value for y_2.

In any event,

$$\frac{t}{s} = \left[\frac{1 + \frac{y_2}{c}}{1 + \frac{1}{c}}\right]^{-k} = \left[\frac{c + y_2}{c + 1}\right]^{-k} \to y_2^{-k} \qquad (\text{as } c \to 0).$$

Thus, as p ranges from 0 toward $-\infty$, the limiting value of t/s ranges from 1 to 0 and we see that we have filled up only one octant.

Moreover, if we calculate coordinates for y^p in what seems to be the same way, namely,

$$s = 1 - (\text{value at } 1) = 1 - 1^p = 0,$$

$$t = 1 - (\text{value at } y_2) = 1 - y_2^p \approx -\, p \log y_2,$$

we find that

$$\frac{t}{s} = -\infty.$$

Either we must accept a slit of some sort between $\{(y + c)^p,\ c \to 0\}$ and $\{y^p,\ p \to 0\}$ or we must recognize some difficulty with our procedure.

Actually, closer examination shows us that the coordinate system we are using does not work at all well for y^p. It is set up to work for $1 - c^k(y + c)^p$, which is 0 for $y = 0$, and approaches 1 from below for all $y > 0$, most slowly for the smallest y. On the other hand, y^p is 0 for $y = 0$ and approaches 1 from *above* for all $y \geqq 1$, most slowly for the largest value of y. These are not compatible.

Essentially the only renormalization we can try is to expand $y^p(1 - Cp)$, which leads to

$$\frac{t}{s} = \frac{C - \log y_2}{C},$$

which can never be made to vanish for more than one value of y_2.

We clearly should give up any attempt to get second-order accuracy in our representation around the far corner. Indeed, when we reflect on the effects of varying values of y_2, we see that we are in the situation discussed in Section 6. Not only second-order accuracy, but also first-order accuracy should be given up. We should arrange to have these curves intersect on the chart, even though they are tangent in the family. The far corner is really a singularity!

23. The far corner if $y + c$ is always positive. In view of the essential way in which the far singularity depended on $y + c$ either being actually zero or approaching zero, and in view of the fact that we may often have situations where $y + c$ will not approach zero, it is probably worth while to reconsider our nor-

malization and to look separately at the case where $y + c$ is always positive. It will be convenient to normalize so that $y \geqq 1$, $c \geqq 0$, as we may clearly do.

The only singularity is now in the vicinity of $p = -\infty$, where we have

$$1 - (1 + c)^k(y + c)^p \approx \operatorname{sgn}^+(y - 1) - \left(\frac{1 + c}{y + c}\right)^k \operatorname{sgn}^+(y - 1) \qquad (c \geqq 0)$$

$$1 - \frac{e^{my}}{e^m} \approx \operatorname{sgn}^+(y - 1) - e^{m(y-1)} \operatorname{sgn}^+(y - 1) \qquad (m < 0)$$

and where, therefore, if we adopt as coordinates s and t the deviations of the values from $\operatorname{sgn}^+(y - 1)$ at the next lowest values y_2 and y_3 of y, we find

$$s = \left(\frac{1 + c}{y_2 + c}\right)^k = e^{m(y_2-1)},$$

$$t = \left(\frac{1 + c}{y_3 + c}\right)^k = e^{m(y_3-1)},$$

and we see that the limiting value of t/s is always zero. Hence all the curves for $p \to -\infty$ (and the curve for $m \to \infty$) are tangent. Moreover the final approach to $\operatorname{sgn}^+(y - 1)$ is exponential in speed.

For practical charting, we may either bring all the curves to a common tangent or leave them separate, so long as we pay some attention to the curves $s =$ constant. If we let the curves intersect at finite angles, quantities which depend continuously on the transformation will be continuous on the chart—they will merely tend to be all more and more closely constant on the curves $s = s_0$ as $s_0 \to 0$.

The exact nature of the curves $s = s_0$, as well as the exact nature of the tangency depends on the value of y_2. If we adopted a chart with tangency, we would have to have a separate one for each value of y_2 to avoid introducing discontinuities. If we adopt a chart without tangency, we may hope that the curves $s = s_0$ will not be too badly distorted over a range of values for y_2. Thus it again seems better to use a non-tangent chart.

If y_2 is arbitrarily close to 1, say $y_2 = 1 + \epsilon$, we find

$$s = \left(\frac{1 + c}{y_2 + c}\right)^k = \left(\frac{1}{1 + \dfrac{\epsilon}{1 + c}}\right)^k \approx \exp\left[-k(\epsilon/1 + c)\right],$$

so that $-k/(1 + c) = p/1 + c$ is a natural parameter (which is constant on the curve which is the limits of the curves $s = s_0$ as $y_2 \to 1$). We find that m is equivalent to $p/1 + c$. Thus it may be well to keep the curves

$$\frac{p}{1 + c} = \text{constant}$$

relatively short as $p \to -\infty$.

24. Conclusions. We thus conclude that we need two forms of chart, depending on whether $y + c = 0$ is to be considered reasonable or not. Except near $\text{sgn}^+ y$, both charts are to follow the topology of Fig. 5. Near the identity transformation, both charts should conform to Fig. 6.

Near $\text{sgn}^+ y$, both charts should distort the topology of the family by opening out the angle and making the previously tangent curves meet at finite angles. The details of these openings out are not clearly prescribed. In the case where $y + c$ is safely > 0 we will probably wish to keep the curves $p/(1 + c) = \text{constant}$ fairly short. The most obvious difference in the two cases will be that if $y + c$ can vanish, then $(y + c)^p$ approaches $\text{sgn}^+(y + c)$ as $p \to 0$; while if $y + c$ can not vanish, it approaches $\text{sgn}^+(y - y_{\min})$, but only as $p \to -\infty$.

VI. Charting the Simple Family

25. The general case (where $y + c$ can vanish). If we are to be fully prepared for the case where $y + c$ can even be zero, then we will want our chart to have the following characteristics:

(A) In the vicinity of y itself, the topology and differential properties should be those suggested by the analysis of Section 21:
  (1) the points with $0 \leqq c \leqq \infty$, $p \leqq 1$ fill up the vertex of a quadrant; and
  (2) polar coordinates are, roughly, given by $\tan \theta = c$, $r = (1 - p)/\sqrt{1 + c^2}$.

(B) Along the arc $c = 0$, $1 > p > 0$ the spacing of the values should be, roughly, uniform in p as suggested by Section 8.

(C) In the vicinity of $\text{sgn}^+ y$ the topology should either be as suggested by Fig. 5 (and Section 20) or the tangency of the curves with $c = \text{constant}$, $p \to -\infty$ should be replaced by intersection. (See Section 23.)

(D) Along the arc $c = 0$, $1 > p > 0$, the curves $p = \text{constant}$ should intersect the arc and not be tangent to it. (See Section 17.)

We have, then, to find a chart form with these properties.

To begin with, we have a figure composed of the vertex of a quadrant, and two curves running out the sides of the quadrant which eventually meet again (at $\text{sgn}^+ y$). The simplest figure which we know with these properties is a quarter sphere, stretching from pole to pole. In the vicinity of each pole we have a quadrant, and the meridians which bound this quadrant meet at the opposite pole. Thus, as a first step, we agree to try taking y as one pole and $\text{sgn}^+ y$ as the other. The arc$\{y^p\}$ becomes a meridian, the arc$\{e^{mv}\}$ becomes a meridian 90° removed from this, and the central region fills in the quarter sphere between the two.

Let θ be an angle describing the meridians, so chosen that $\theta = 0°$ for the arc $\{y^p\}$ which has $c = 0$, and $\theta = 90°$ for the arc $\{e^{mv}\}$ which corresponds to $c = \infty$. Let ψ be an angle describing the parallels of latitude, with $\psi = 0°$ for the pole representing y, $\psi = 90°$ at the equator, and $\psi = 180°$ at the pole representing $\text{sgn}^+ y$. Our problem is to find analytic representations for θ and ψ in terms of p and c which will give us the desired properties. These we find by successive approximations.

Near the y pole, we want

$$\tan\theta \sim c, \qquad \psi \sim \frac{1-p}{\sqrt{1+c^2}},$$

and when we observe that $1-p$ ranges up to infinity, it is natural to try

$$\tan\theta = c, \qquad \tan\frac{\psi}{2} = \frac{1-p}{\sqrt{1+c^2}}.$$

But when we realize that while $1-p$ ranges up to $+\infty$ for $c \neq 0$, it only ranges up to 1 for $c = 0$, it is natural to try, perhaps,

$$\tan\theta = c, \qquad \operatorname{ctn}\frac{\psi}{2} = \frac{c}{1-p} + \frac{1}{1+c}\tan^+\frac{\pi p}{2},$$

where we have (i) replaced $\sqrt{1+c^2}$ by c in order to have the first term ineffective for $c = 0$; (ii) introduced the new term with a factor $1/(1+c)$ to make it ineffective for $c = \infty$, where the first term $\sim -c/p$, which is appropriate; (iii) gone to the cotangent instead of the tangent in order to make the combination of the two terms easy; and (iv) introduced $\tan^+\pi p/2$ as a function spreading the values of p from 0 to 1 out uniformly along the arc $c = 0$.

This choice of functions gives a rather reasonable chart, but it still fails to meet our requirements in that (i) the curves $\{(y+c)^p,\ c \to 0\}$ are tangent to $\{y^p\}$, and (ii) near $\operatorname{sgn}^+ y$ we have not begun to obtain the proper local structure, since the curves $c =$ constant intersect, while the curves $p =$ constant are mutually tangent. We can try to get at both difficulties at once by changing to

$$\tan\theta = c + (1-p)\sqrt{c}, \qquad \operatorname{ctn}\frac{\psi}{2} = \frac{c}{1-p} + \frac{1}{1+c}\tan^+\frac{\pi p}{2}.$$

This change pushes the curves $c =$ constant upward, more strongly for lower c and larger $1-p$, and thus tends to meet both needs. However, the curves $p =$ constant are still mutually tangent to the boundary curve $c = 0$ at $\operatorname{sgn}^+ y$. In order to correct this in a simple way, we need only add a $(-p)^+$ term, which will vanish for $p \geqq 0$. The result is

$$\tan\theta = c + (1-p)\sqrt{c} + (-p)^+, \qquad \operatorname{ctn}\frac{\psi}{2} = \frac{c}{1-p} + \frac{1}{1+c}\tan^+\frac{\pi p}{2}.$$

26. Plotting. We now have represented the central part of the simple family on a quarter-sphere. For practical purposes we require a plot on the plane. The solution adopted here was influenced substantially by the existence of some special equal-area graph paper which the writer obtained from Mount Wilson Observatory during World War II in connection with problems of mutual interest. This graph paper represents a half-sphere area-true on a rectangle and provides two families of spherico-polar coordinates, one with poles at the top and bottom of the rectangle and the other with poles at the centers of the right and left sides. If u, r are rectangular coordinates with $0 \leqq u \leqq \pi$, $-1 \leqq r \leqq +1$

defining the rectangle, then ψ and θ are related by

$$\cos \psi = \sqrt{1 - r^2} \cos u,$$

$$\cos \theta = \sqrt{1 - r^2} \csc \psi,$$

$$\operatorname{sgn} \theta = \operatorname{sgn} r.$$

The combination of this representation of the half-sphere on the plane with the map on the quarter-sphere gives the plot already presented in Fig. 1.

27. The case of moderate $y + c$. In case $y + c$ cannot vanish, we want a plot with a different and simpler topology. The choice

$$\tan \theta = c,$$

$$\tan \frac{\psi}{2} = \frac{q}{\sqrt{1 + c^2}},$$

where the transformation is written in the form

$$(y + cy_0)^{1-qe_0},$$

scaled as to size by y_0 and as to change in exponent by e_0, is the natural transformation near $q = 0$, and meets our requirements everywhere else.

When combined with the same graph paper, the result is as in Fig. 3.

REFERENCES

[1] F. J. Anscombe and J. W. Tukey, "The criticism of transformations," paper presented before the American Statistical Association and the Biometric Society, Montreal, September 12, 1954.

[2] P. G. Moore and J. W. Tukey, "Answer to query 112," *Biometrics*, Vol. 10 (1954), pp. 562–568.

THE PROBLEM OF ESTIMATION

By H. Steinhaus

University of Wroclaw

0. Introduction. A treatment has proved successful in m cases out of n; what is its efficacy? In other words: What will be the frequency of successes if we continue to apply the same treatment in all cases of the disease it was invented for? A lot has been examined by sampling and m items out of n proved defective; what is the best estimate for the fraction defective in the whole lot? This is another example of the same question. A hundred people of a certain tribe have been classified as to their belonging to the blood groups A, B, AB and O; how are we to estimate the frequencies of the 4 groups in the whole tribe? It is not an overstatement to say that the abundance of applications gives the problem of estimation an importance sufficient to rank it among the principal problems of science. This is the reason why, for more than a century, it has not ceased to haunt the ingenuity of mathematicians.

To make things quite plain let us imagine a statistician (S) compelled by the devil (D) to play the following game:

The devil has a collection of coins and he knows for each of them its probability p of showing heads as a result of tossing; he is rich enough to have specimens for any p in (0, 1). He chooses a coin suiting his fancy and lets the statistician throw it n times; it shows heads m times and it is up to the statistician to give an estimate p' of the value p which is known to D but unknown to him. This being done, S pays to D \$$(p' - p)^2$. S tries his best to reduce his loss by an appropriate method of guessing, as far as possible. If he succeeds in finding the best method he will not regret the money lost: his will be the fame of having solved our problem of the best estimate, if "best" is understood as minimizing the expected square error of the estimation; the rules of the game have been fixed by the devil in accordance with such an interpretation of our problem.

The following remark may explain the link connecting our game with the problem of point-estimation. The classical method solves this problem assuming that the distribution of coins employed by the devil is known to the statistician. He computes his guess, combining by Bayes' rule this knowledge with the observed result of tossing; the guess p' will be equal to the a posteriori value Ep. It can be defined also as the value of p' that minimizes the expected loss $E(p' - p)^2$. Thus the rules of our game correspond to the problem of point-estimation in the case of a known prior distribution; it is not artificial to employ them in the case of an unknown prior distribution.

I proposed this problem in 1954 [10] in Prague and in Berlin, calling it "Das statistische Spiel," but it is only in 1956 that I have been informed by L. J. Savage that it has already been solved: J. L. Hodges, Jr., and E. L. Lehmann [4]

Received July 30, 1956, revised April 22, 1957.

attribute the priority to Herman Rubin in their important paper published in 1950. Another paper [3] by M. A. Girshick and L. J. Savage deals thoroughly with the same question; the results have already been included in books [2] and [8]. The generality of the results reached by the American school and their priority make it necessary to explain the reasons for publishing this paper: the first is that my exposition will perhaps be easier to read for some readers who do not feel comfortable when spoken to in the language of consummate specialists; the second is to be found in the last sections, which deal with some cases not covered by the authors quoted above.

1. A statistician who does not believe in probabilities when he is faced with a single game does not even look at the gambling table. He says: "The number m gives me no information whatever about p. My method is to declare p equal to $\frac{1}{2}$. In the worst case the true p will be 0 or 1 and I will lose 25 cents. Any gambler adopting a system different from mine exposes himself to a greater loss than a quarter of a dollar: if his p' is greater than $\frac{1}{2}$ he loses more than 25 cents for a very small p and if his p' surpasses $\frac{1}{2}$ his loss is greater than 25 cents for a p near unity. This is the reason why I abstain from counting heads and tails."

The method described above denies the possibility of statistical estimation, since its estimate is not influenced by the results of experiments. It has, however, the advantage of the certainty it gives to S of never losing more than 25 cents, a guarantee that no other method is capable of. From this point of view we can define 25 cents as the "value of the game." The devil's point of view, however, is different. As he is never sure that S will not guess the exact p, in which case his gain is nil, the "value of the game" is 0 when computed by D. In both cases we consider only the gains of D; nevertheless we get two different values: 25 cents and 0. Such games may be called "open."

2. As is generally known, many mathematical theories were invented to master our problem by the calculus of probability. Laplace did not hesitate to apply to it the so-called inverse probabilities or "probability of causes" derived from the observation of effects. To avoid misunderstandings we will speak of frequencies of different coins instead of the probabilities of their having been chosen by D. Such an approach implies the substitution of a sequence of games for a single game, and of the mean loss for the single loss. Following this principle, we have to imagine S being condemned by D to perpetual gambling. The devil is allowed to choose a new coin in each round of the infinite sequence Q of rounds, but n remains unaltered. It is the statistician's business to make his mean loss as small as possible. He could do it easily if he knew the frequency of different coins employed. To be more explicit, let us assume that D adheres to a certain function $f(p)$, the "density," and that S knows this function. In other words, he knows that for any two numbers u, v $(0 \leqq u \leqq v \leqq 1)$ the relative frequency of the coins with $u \leqq p \leqq v$ is $\int_u^v f(p)\,dp$. The function $f(p)$ is integrable, nonnegative and $\int_0^1 f(p)\,dp = 1$. S puts the following problem to himself: the whole sequence Q of games can be split up into $n + 1$ subsequences Q_m $(m = 0, 1, 2, \cdots, n)$,

Q_m being the subsequence consisting of all rounds showing the same number m of heads. The mean loss in such a subsequence is proportional to the integral

$$(1) \qquad I(x, m) = \int_0^1 f(p) \binom{n}{m} p^m q^{n-m} (p - x)^2 \, dp \qquad (q = 1 - p),$$

where x is the statistician's answer to m heads. Thus he has to find a function $p' = g(m, n)$ which gives to $I(x, m)$ the least value when substituted for x in the integral (1). Since the $(n + 1)$ values p' takes for $m = 0, 1, \cdots, n$ are independent, S has to solve the problem $I(x, m) =$ minimum separately for each m. This leads to a function $g(m, n)$ that minimizes the mean loss in the whole sequence Q. The problem is easily solved, for x is the root of the equation $dI(x, m) / dx = 0$. Calling it x_m we get immediately

$$(2) \qquad x_m = \int_0^1 f(p) p^{m+1} q^{n-m} \, dp \Big/ \int_0^1 f(p) p^m q^{n-m} \, dp = g(m, n).$$

Formula (2) is the clue to the problem: S has to declare $p' = g(m, n)$ in every game, $g(m, n)$ being defined by the ratio (2). It is to be noted that the integral (1) is a nonnegative quadratic function of x, which implies that the root of $I' = 0$ is a perfect minimum of I: for any x different from x_m, we get $I(x, m) > I(x_m, m)$.

A nontrivial case is given by the Bayes hypothesis $f(p) \equiv 1$. Formula (2) then leads immediately to

$$(3) \qquad g(m, n) = \frac{m + 1}{n + 2}$$

as the best method for S. Laplace employed such a hypothesis and (3), and he has been rebuked for it by his successors. An example of another kind is the favorite method of many naturalists and other scientists, who put $p' = m/n$. We now have to determine $f(p)$ so as to satisfy (2) identically in m (n is fixed once for all) with $g(m, n)$ replaced by m/n. Putting $m = n$ we immediately get the condition

$$(4) \qquad \int_0^1 f(p) p^{n+1} \, dp = \int_0^1 f(p) p^n \, dp.$$

The functions integrated being nonnegative and p^{n+1} less than p^n everywhere in (0, 1), (4) implies $f(p) \equiv 0$ (almost everywhere), which contradicts the condition $\int_0^1 f(p) \, dp = 1$, essential for any density. Thus the estimate m/n does not correspond to any density $f(p)$ whatever. It is a remarkable fact that some surgeons have reached the same conclusion without being mathematicians and without knowing how to improve the estimate [7].

3. The modern theory of games as codified in [5] and extended, for example, in [12], guarantees the existence of a solution of our problem. It proves, for such two-person games as are considered here, the existence of a density $f(p)$ of a sequence $\{g_i(m, n)\}$ of guessing-methods, and of a number V (called the value of the game) such that

1^0 the statistician applying the guess $p' = g^{(j)}(m, n)$ in the jth round is sure to lose no more than V in the mean whatever the devil does, the mean being taken over the results of the whole sequence Q.

2^0 the devil who mixes his coins in conformity to the frequency function $f(p)$ is sure to win in the mean at least V regardless of what S does. The use of the frequency function $f(p)$ by D is to be understood as the use of a sequence of coins such that the corresponding sequence $\{p^{(j)}\}$ has a relative frequency $\int_u^v f(p)\,dp$ of terms belonging to the interval (u, v) whatever this interval may be $(0 \leqq u \leqq v \leqq 1)$. It follows immediately from 1^0 and 2^0 that

3^0 the "mixed strategy" defined by $\{g^{(j)}\}$ is the best for S in the minimax sense of "best."

It is also clear that V, the "value of the game," is the fixed indemnity which, when paid by D to S before every choice of a coin, makes the play fair; such settlement would not influence our theory.

The importance of the theorem of J. von Neumann leading (through its extensions) to the statements mentioned lies, as far as our endeavor is concerned, in the assurance that it is worthwhile to seek an effective formula for the estimate p'. Nevertheless, our reader is neither supposed to know the theory of games nor the somewhat intricate terminology of the books quoted ([5] and [12]). Our argument will be independent of the general theorem, and it will yield in our special problem more than is promised by 1^0, 2^0, and 3^0. The difference between the result of Sec. 1 and the objective of the following sections is also worth noticing: we could not guarantee to S a loss less than 25 cents in a single round but we can—as will be shown—reduce his mean loss in a sequence of rounds to a quantity V less than 25 cents by finding a minimax solution which admits no further improvement.

It is necessary to explain why we are not entirely satisfied by the promise contained in 1^0, 2^0, and 3^0. What we need is an estimate $p' = g(m, n)$ valid for the whole sequence Q of successive rounds; estimates varying from round to round would introduce an indeterminism of a dangerous kind which would make the scientist utter different judgments in identical situations: if $g_j(m, n) \neq g_h(m, n)$ for the same couple m, n, he is compelled to estimate differently the efficacy of two antibiotics, although both of them gave m positive results in n experiments; he is compelled to such behavior because the first experiment was scheduled "n_0, j" and the other "n_0, h." In Sec. 1 we have seen that no satisfactory estimate can be reached by making it an absolute constant; now we reject estimates which have the drawback of variability, though it is not universally agreed that such estimates are irrational. The case against them is forcibly put by R. A. Fisher, pp. 97–98 of [1].

4. Let us suppose now that D has chosen once for all a particular coin characterized by p, whereas S employs the particular function $g(m, n)$ of (3) as his method of guessing, and let us call E the mean gain of the devil. We get

$$(5)\quad E = \sum_{m=0}^{n} \binom{n}{m} p^m q^{n-m} \left(p - \frac{m+1}{n+2}\right)^2$$

$$= \sum_{m=0}^{n} \binom{n}{m} \left[p^2 - 2p\frac{m+1}{n+2} + \left(\frac{m+1}{n+2}\right)^2\right].$$

To compute E we make use of the trivial equality $\sum_{m=0}^{n} \binom{n}{m} p^m q^{n-m} = 1$ and of the following identities:

$$(6)\qquad \sum_{m=0}^{n} \binom{n}{m} m p^m q^{n-m} \equiv np, \qquad \sum_{m=0}^{n} \binom{n}{m} m^2 p^m q^{n-m} \equiv n^2p^2 + npq;$$

they are, respectively, familiar formulae for the first and second moments of the binomial distribution. Eqs. (5) and (6) lead immediately to the simple result

$$E = \frac{(n-4)pq + 1}{(n+2)^2}.$$

Let us consider first the case $n = 4$. Then E ceases to depend on p and becomes equal to 1/36. This circumstance makes the statistician who adopted the method $p' = (m + 1)/6$ independent of all diabolic cunning: whatever D does, whether he sticks to one coin or changes them as he likes, S will lose in the long run 1/36 as his mean expenditure, never more nor less. It follows therefrom that any method is the best for D as an answer to the method $p' = (m + 1)/6$. In particular, the method defined by the density $f(p) = 1$ is such a relatively best method against the statistician's rule $p' = (m + 1)/6$. What is, on the other hand, the statistician's adequate answer to the devil's rule $f(p) \equiv 1$? This answer has already been found in Sec. 2: $g(m, n) = (m + 1)/(n + 2)$, which is equal in our special case $n = 4$ to $(m + 1)/6$. Thus we have two antagonistic methods for D and S, each of them being relatively best against the other. We can easily prove that the rule of guessing $p' = (m + 1)/6$ is not only relatively best but also absolutely best, i.e., that it fulfils the minimax criteria for best strategies. The proof runs as follows: The rule $p' = (m + 1)/6$ reduces the loss of S to 1/36 whatever D does. On the other hand, the same rule is the best answer against D's employing the system $f(p) \equiv 1$. This last statement implies the impossibility of S's reducing his loss below 1/36 against the devil's playing as mentioned. Therefore no method gives a certainty to S of a loss below 1/36; as the rule above gives a guarantee of a loss equal to 1/36, no rule is better in the minimax sense, Q.E.D. The devil's rule $f(p) \equiv 1$ is also absolutely best for D, a theorem which can also be proved almost immediately. Both proofs are superfluous for a reader knowing the general theorem about games that two methods relatively best against each other are also absolutely best for the opponents employing them. The existence of such a pair of methods shows, in our terminology, that the game is closed. The result of their being constantly applied by both sides is a mean loss equal to the value of the game, which is 1/36 in our case.

Thus we have solved our problem of the best estimate for $n = 4$. The solution is $p' = (m + 1)/6$.

Let us try to free ourselves from the restriction $n = 4$ by setting more generally,

$$p' = g(m, n) = \frac{m + a}{n + b} \tag{7}$$

and assuming the independence of a and b from m but not from n. To compute E in this case we have to write

$$E = \sum_{m=0}^{n} \binom{n}{m} p^m q^{n-m} \left(p - \frac{m + a}{n + b}\right)^2 \tag{8}$$

instead of (5), and to repeat almost verbally the trivial computations of the special case $a = 1$, $b = 2$; they give now

$$E = \frac{npq + (bp - a)^2}{(n + b)^2} \qquad (q = 1 - p). \tag{9}$$

The numerator $np(1 - p) + (bp - a)^2$ ceases to depend on p just if we put $a = \sqrt{n}/2$, $b = \sqrt{n}$. These values when substituted in (7) for a and b yield the formula

$$p' = g(m, n) = \frac{m + \sqrt{n}/2}{n + \sqrt{n}} = G(m, n) \tag{10}$$

as a guessing method for S. When substituted in (9), they give

$$\Lambda = \frac{1}{4(\sqrt{n} + 1)^2} \tag{11}$$

as the mean loss of the statistician, whatever the sequence of coins employed by the devil. To follow the path traced in the special case $n = 4$, we have to find a density $f(p)$ which satisfies (2) when we write for $g(m, n)$ the function G defined by (10). Let us try to this effect

$$f(p) \equiv c(pq)^s (s > -1) \quad \text{with} \quad 1/c = \int_0^1 (pq)^s \, dp.$$

The integrals in (2) now take the form

$$c \int_0^1 p^{s+m+1}(1 - p)^{s+n-m} \, dp, \qquad c \int_0^1 p^{s+m}(1 - p)^{s+n-m} \, dp,$$

and their respective values are

$$\frac{\Gamma(s + m + 2)\,\Gamma(s + n - m + 1)}{\Gamma(2s + n + 3)}, \qquad \frac{\Gamma(s + m + 1)\,\Gamma(s + n - m + 1)}{\Gamma(2s + n + 2)}.$$

Their ratio is $(s + m + 1)/(2s + n + 2)$, and condition (2) becomes

$$\frac{m + s + 1}{n + 2s + 2} = \frac{m + \sqrt{n}/2}{n + \sqrt{n}}. \tag{12}$$

To satisfy (12), we set $s = \sqrt{n}/2 - 1$ and get

$$f(p) \equiv (pq)^{\sqrt{n}/2-1} \cdot \frac{\Gamma(\sqrt{n})}{(\Gamma(\sqrt{n}/2))^2} \tag{13}$$

as the density sought to close the game. Thus the method G defined by (10) is the solution of our problem. It is the best estimate of p because it minimizes the mean square error. The case $n = 4$ is obviously contained as a special case in (10); it is the only case in which the hypothesis of uniform distribution a priori is consistent, in a certain sense at least, with our theory.

It is interesting to note that the classical rules—Bayes' hypothesis of uniform prior distribution and the estimate m/n—can both be saved if $(p - p')^2/pq$ is the amount to be paid by S to D. This is easily verified by writing $a = 0, b = 0$ in (7) and dividing (9) by pq to compute the expected gain of D in the new game; it is $1/n$ and thus independent of p. On the other hand, if we put $f(p) \equiv 1$ in (1) and replace $(p - x)^2$ by $(p - x)^2/pq$ to satisfy the new rules of the game, we are led by the same reasoning as employed for the old rule to the formula

$$x_m = \int_0^1 p^m q^{n-m-1}\, dp \Big/ \int_0^1 p^{m-1} q^{n-m-1}\, dp = m/n,$$

instead of (2).

5. The method of S given by G holds the devil tightly by his horns: he can neither increase nor diminish his mean gain which is equal to Λ as defined by (11). Λ is the value of the game and, consequently, the indemnity which, when paid in advance by D to S in every round, makes the whole game fair: this signifies that the mean balance of the sequence of rounds tends to zero with probability 1 if this indemnity is agreed upon and both partners play as well as possible. If S adheres to G the devil's lack of skill does not alter the limit. There are, however, bad methods for D (i.e., bad densities $f(p)$); if S knew that D has chosen such a method, he could apply a method securing himself a positive mean balance.

The expression

$$\sqrt{\Lambda} = \tfrac{1}{2}(\sqrt{n} + 1) \tag{14}$$

may be called the mean error of the estimate (10). We emphasize that it is dependent exclusively on n. This enables the scientist to plan his experiments with a desired accuracy: he gets by an adequate choice of n the accuracy desired, neither more nor less, which is very important (e.g., the estimation of a lot by counting good and defective items in a sample of n has to avoid samples too great and too small: the first cost too much, the second give insufficient information). To appreciate this feature of the estimate (10) we may compare it with the "confidence interval" method where the length of the interval, given the confidence level, is known only after the experiment, because it depends on both n *and* m. Finally, our formulae are extremely simple, and can easily be tabulated for practical purposes.

6. Our estimate G is free from the disadvantage of "mixed strategies" anticipated at the end of Sec. 3. We may put the question of its uniqueness. Stanisław Trybuła has found the following simple proof [11] that there is no other estimate *equivalent* to G. It has already been proved here that there is no *better* estimate; we have the author's consent to repeat here his proof of uniqueness. The proof that there is no other estimate *equivalent* to G is simple. It is stated, in more general form in [4], where the estimate was first announced. I happened to hear it first from Stanisław Trybuła who rediscovered it in connection with [11]. You have already seen that there is no better *estimate*; and now here is the proof of uniqueness.

We have already stated in Sec. 2 that the solution (2) of the minimum problem for the integral $I(x, m)$ defined in (1) is unique:

$$I(x, m) > I(x_m, m) \text{ for every } x \text{ different from } x_m . \tag{15}$$

Suppose that S adopts any method M, of the form $g(m, n)$ different from G. Let the devil choose the method defined by (13). Since this density has been shown to give G when substituted in (2), we see by (15) that there is at least one subsequence G_m giving to D a greater gain in the mean than the gain he would obtain if S applied the method M: this subsequence is generated by the subscript m for which M differs from $G(m, n)$; there is at least one such value m among $0, 1, \cdots, n$. For any values m, for which M equals $G(m, n)$ the corresponding subsequences G_m give the same gain to D as he would gather if opposed by $G(m, n)$. Thus S will suffer in the mean over all rounds played a loss greater than L, if he applies M and is opposed by (13) as the devil's method. Thus M is not a minimax solution, which proves the uniqueness of our minimax solution $G(m, n)$.

The question of uniqueness of the devil's best method has a different answer. It too has been found by Trybuła [11], and it is negative. There are densities different from (13) that are equivalent to it as gambling methods in every respect. They are, however, not so simple.

We have already seen that there does not exist a single coin p_0 as a "rigid method" for D; the reason is the possibility of S's choosing $p' = p_0$ if D's method is to always use the same coin p_0. Nevertheless, there are other methods, which correspond neither to densities nor to constants, to speak more generally, which have no distributions. We have omitted taking them into account; because our argument led first to a method G for S which made his mean loss independent of all methods of D, the distributionless methods included, and we had only to find one definite method for D which could never be better answered than by G. Thus we can admit all possible methods of D, having a distribution or not; the theory expounded remains valid without restrictions. As the devil represents nature, we have no reasons to jeopardize his freedom; it is an advantage for S, the human player, that our theory enables him to master the situation in the *general* case.

In all cases, when the existence of limits is asserted, the assertions are true with probability one. There are other points where we have sinned against the

professional vocabulary, e.g., when speaking about the "mean error"; it is rather a metaphor: we always think of the expression

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} (p^{(j)} - p'^{(j)})^2.$$

For practical purposes this limit is a better measure for the error than any having the disadvantage of being dependent on m and of becoming 0 for $m = 0$ and for $m = n$. These disadvantages have shocked practicians and compelled them to leave the hard soil of sound statistics ([7], [9]).

Our method may be criticized as leading to a biased estimate. This means, in plain words, that in the case of estimating the same quality p over and over, say r times, we get r values m_j $(j = 1, \cdots, r)$ and r estimates

$$p'^{(j)} = \frac{m_j + \sqrt{n}/2}{n + \sqrt{n}}$$

whose mean $1/r \sum_{j=1}^{r} p'^{(j)}$ does not tend to p for $r \to \infty$. This is, however, an artificial situation. If we have different lots L_j we are not entitled to suppose $p^{(j)}$ the same for all lots and, to be on the safe side, we must give separate estimates for each lot. If, however, we have to deal with the same lot r times, we will not compute its quality by taking the mean of the estimates but we will employ the formula (10) with $m = \sum_{j=1}^{r} m_j$, writing

$$p'_r = \frac{m + \sqrt{nr}/2}{nr + \sqrt{nr}}.$$

The estimate p'_r has, obviously, the property that $\lim_{r\to\infty} p'_r = p$.

Our method is essentially based on the measure of the error given by $(p - p')^2$. Of course, this measure is by no means the only possible one. In some applications the minimizing of the average error $| p - p' |$ would be more realistic; in some cases even some asymmetric function of the difference $p - p'$ could claim to be the best measure of the error in view of the economic consequences of a bad estimate which ought to be minimized. Notwithstanding this objection, we do not see any possibility of covering all the statistical applications by a single formula and must either abstain from giving any general rule or adopt the principle of least squares as being by far the best known and most applied. Its simplicity is inherited by our formula; no tables are necessary to employ it for small n, and there is no distinction between small and large n to be taken care of.

7. The comparison of our estimate (10) and that called the "confidence interval" [6] leads to the following remarks.

The confidence interval is defined by a "confidence level" and by the postulate of being the shortest one, given m, n and the confidence level. It is not a direct estimate of the unknown p. In many cases, however, such a direct estimate is necessary; let us think, for instance, of selling two kinds of some produce on the basis of a first experience which gave m and $n - m$ buyers respectively for kinds

A and B; the producer is compelled to choose a definite p' so that he may include in the next shipment the proper numbers of items A and B. Nevertheless we could consider our estimate as the middle point of an interval of the length $1 / (\sqrt{n} + 1)$. This length being the double of the expression (14), we may risk the conjecture that the confidence level for it is $0.6827 \cdots$ for n greater than, let us guess, 30. Thus our estimate would appear as a special case of the classical confidence interval: the case of confidence level 0.6827. This is, however, an illusion, because of the classical confidence interval's having a length depending on m which is not true for the length (14). We have already pointed out in Sec. 5 the importance of this independence.

The remarks above lead to the problem of modifying the so-called confidence interval by making its length independent of m. To speak again in the language of the theory of games, it would be necessary to let the devil and the statistician fix n and an arbitrary t ($0 \leqq t \leqq 1$) and let them gamble, the devil choosing the coin to be tossed and the statistician placing an interval J of length t on the real axis after having seen the result of n tossings. The stake would be \$1 paid by D to S if p is covered by J and by S to D in the opposite case. The problem for S would consist in placing the middle-point P of J as well as possible. He would have to find a formula $P(m, n; t)$ for the best location of P. This formula would be the statistician's best method of playing and would lead eventually to a value of the game, $V(n, t)$, independent of p. Solving the equation $\frac{1}{2}V(n, t) = v$ we would get $n = h(t, v)$; the function $h(t, v)$ would answer the following problem: given the confidence level v and the length t, to determine the number n of experiments needed to include the unknown p in an interval of length t with a confidence v. We have not tried to solve this problem; we mention it here to show distinctly the difference between the theory of Neyman's confidence intervals and our line of approach.

We pass now to another question; it is the last of the three examples spoken of in the introduction.

Instead of the simple alternative, or dichotomy, for which the solution has already been described, we have here four possible results of tossing. Let us more generally suppose that there are k possible outcomes, $k > 2$. The coin must be replaced by a k-hedron, every face being painted differently from the others; D knows the respective probabilities $p_i (i = 1, 2, \cdots, k)$ for all faces; he has in his collection a model for every set $\{p_i\}$; obviously $\sum_{i=1}^{k} p_i = 1$. The result of n tossings with a definite k-hedron is a set $\{m_i\}$; m_i $(i = 1, 2, \cdots, k)$ are the respective numbers of tossings resulting in the ith face up; obviously $\sum_{i=1}^{k} m_i = n$. S has to estimate $\{p_i\}$ knowing $\{m_i\}$; his estimate $\{p_i\}$ has to minimize the expected value of

$$\sum_{i=1}^{k} (p_i' - p_i)^2. \tag{16}$$

He tries to do it by putting

$$p_i' = \frac{m_i + a}{n + b} \qquad (i = 1, 2, \cdots, k). \tag{17}$$

We now have, instead of (5), the expression

$$E = \sum_{(m)} \frac{n!}{m_1!\, m_2! \cdots m_k!}\, p_1^{m_1} p_2^{m_2} \cdots p_k^{m_k} \cdot \left\{\left(p_1 - \frac{m_1 + a}{n + b}\right)^2 + \cdots + \left(p_k - \frac{m_k + a}{n + b}\right)^2\right\}, \tag{18}$$

the summation being extended over all systems of nonnegative integers m_i with $\sum_{i=1}^{k} m_i = n$. Let us denote by E_1 the part of E which is left after cancelling all squares in $\{\ \}$ in (18) but the first. If we denote m_1 by m, p_1 by p, and $\sum_{i=2}^{k} p_i$ by q we get immediately the same expression for E_1 that we have called E in (8). This gives, in view of (9)

$$E_i = \frac{1}{(n + b)^2} \cdot (np_i(1 - p_i) + (bp_i - a)^2).$$

The same applies to each of the k parts E_i of (18) and leads, eventually, to

$$\begin{aligned} E = \sum_{i=1}^{k} E_i &= \frac{1}{(n + b)^2} \sum_{i=1}^{k} \{np_i(1 - p_i) + (bp_i - a)^2\} \\ &= \frac{1}{(n + b)^2}\left\{n + (b^2 - n) \sum_{i=1}^{k} p_i^2 - 2ab + ka^2\right\}. \end{aligned} \tag{19}$$

This expression becomes independent of the variables p_i if and only if we put $b = \sqrt{n}$; it then takes the form

$$E = \frac{1}{(n + \sqrt{n})^2}(n - 2a\sqrt{n} + ka^2). \tag{20}$$

To determine a we introduce the condition

$$\sum_{i=1}^{k} p_i' = 1, \tag{21}$$

obviously needed in most applications. Thus we get from (17) with the value $b = \sqrt{n}$

$$1 = \sum_{i=1}^{k} \frac{m_i + a}{n + b} = \frac{n + ka}{n + \sqrt{n}}, \tag{22}$$

and, from (22), $a = \sqrt{n}/k$. We note with gratification that this value of a necessary to insure (21) is also just the value needed in (20) to minimize E without regard for (21). The formulae (17) now become

$$p_i' = \frac{m_i + \sqrt{n}/k}{n + \sqrt{n}} \qquad (i = 1, 2, \cdots, k). \tag{23}$$

Putting $a = \sqrt{n}/k$ into (20) we get

$$E = \frac{n}{(n + \sqrt{n})^2} \cdot \frac{k - 1}{k}, \tag{24}$$

an expression independent of $\{p_i\}$.

To attain our objective by pursuing the analogy with the case $k = 2$, we must find a density on the hyperplane

$$\sum_{i=1}^{k} p_i = 1 \quad (0 \leqq p_i \leqq 1;\ i = 1, 2, \cdots, k), \tag{25}$$

for which the estimates (23) minimize the expected value of $\sum_{i=1}^{k} (p_i' - p_i)^2$. Let us try the density $f(p_1\, p_2 \cdots p_k) = C(p_1\, p_2 \cdots p_k)^s$, where the constants C and s will be determined later. Supposing that the statistician knows that the frequencies of k-hedra employed successively by the devil are distributed according to this density (each point on the hyperplane (25) defines a k-hedron and the converse is also true), he will have to determine his guesses $\{x_i\}$, after having observed the result $\{m_i\}$, so as to minimize the expected loss that is given by the positive quadratic form

$$\begin{aligned} &F(x_1, x_2, \cdots, x_k) \\ &= C\frac{n!}{m_1! \cdots m_k!}\int_{p_1+\cdots+p_k=1} \cdots \int (p_1 \cdots p_k)^s p_1^{m_1} \cdots p_k^{m_k} \sum (p_i - x_i)^2\, dw. \end{aligned} \tag{26}$$

The integral in (26) is extended over the domain defined by (25) and dw is the element of the hyperplane (25). The minimum problem is equivalent to the set of equations $\partial F/\partial x_i = 0$ $(i = 1, 2, \cdots, k)$ which are obviously

$$\begin{aligned} &\int_{p_1+\cdots+p_k=1} \cdots \int \\ &\cdot (p_1\, p_2 \cdots p_k)^s p_1^{m_1} p_2^{m_2} \cdots p_k^{m_k}(p_i - x_i)\, dw = 0 \qquad (i = 1, 2, \cdots, k) \end{aligned} \tag{27}$$

and the method of guessing defined by (23) will be proved best in the minimax sense if the set $\{x_i\}$ solving (27) can be identified with the set $\{p_i'\}$ given by (23). This amounts to k conditions of which the first is

$$\frac{\int \cdots \int p_1^{m_1+s+1} p_2^{m_2+s} \cdots p_k^{m_k+s}\, dw}{\int \cdots \int p_1^{m_1+s} p_2^{m_2+s} \cdots p_k^{m_k+s}\, dw} = \frac{m_1 + \sqrt{n}/k}{n + \sqrt{n}}. \tag{28}$$

The numerator L and the denominator M in (28) can be brought into the forms

$$\begin{aligned} L &= \frac{\Gamma(m_1 + s + 2)\Gamma(m_2 + s + 1) \cdots \Gamma(m_k + s + 1)}{\Gamma[n + k(s + 1) + 1]} \\ M &= \frac{\Gamma(m_1 + s + 1)\Gamma(m_2 + s + 1) \cdots \Gamma(m_k + s + 1)}{\Gamma[n + k(s + 1)]} \end{aligned} \tag{29}$$

respectively, if the classical identity [13]

$$\iint \cdots \int \prod_{i=1}^{k} p_i^{\alpha_i - 1}\, dw = \frac{\prod \Gamma(\alpha_i)}{\Gamma(\sum \alpha_i)} \tag{30}$$

is applied to both of them. (29) immediately yields

$$\frac{L}{M} = \frac{m_1 + s + 1}{n + k(s+1)} \tag{31}$$

and the condition (28) leads, through (31), to

$$\frac{m_1 + s + 1}{n + k(s+1)} = \frac{m_1 + \sqrt{n}/k}{n + \sqrt{n}}, \tag{32}$$

which can be satisfied by setting $s + 1 = \sqrt{n}/k$; since the same reasoning applies to all k conditions of which (28) is the first, all will be satisfied by the value chosen for s. As to the constant C it can now be found from the condition

$$C \int_{p_i \geqq 0; p_1 + \cdots + p_k = 1} \cdots \int (p_1 p_2 \cdots p_k)^{\sqrt{n}/k - 1}\, dw = 1 \qquad (k > 2) \tag{33}$$

which gives

$$C = \frac{\Gamma(\sqrt{n})}{\Gamma^k(\sqrt{n}/k)} \tag{34}$$

and, finally,

$$\frac{\Gamma(\sqrt{n})}{\Gamma^k(\sqrt{n}/k)} (p_1 p_2 \cdots p_k)^{\sqrt{n}/k - 1} \tag{35}$$

is the density sought. If it is employed by the devil in an infinite game where the payment (16) in every round is agreed upon, the statistician's best answer is defined by (23), whether or not S is bound to respect the condition (21). This statement has just been proved; it enables us to apply once more the argument employed in the case $k = 2$ and to recognize the estimates (24) as the best in the minimax sense. The uniqueness of our solution follows by the reasoning already applied (for $k = 2$) in Sec. 6: as F in (26) shares with I in (1) the property of attaining its minimum at a single point, no modifications of the former reasoning are necessary.

8. If we drop the condition (21) we can write p for p_1, q for $1 - p$ and seek with S the best estimate of p_1 alone by the theory of the dichotomous case. This leads immediately to the solution (10) which may be written adequately

$$p_1' = \frac{(m_1 + \sqrt{n}/2)}{(n + \sqrt{n})};$$

we can, of course, replace the subscript 1 on both sides by $i (i = 1, 2, \cdots, k)$. We will have for the total loss (16) the mean $k\Delta$, where Δ is given by (11). On the other hand, the loss has been computed in (24) under condition (21) as being equal to $4[(k-1)/k]\Delta$; since $4[(k-1)/k]$ is less than k for $k > 2$, we arrive at the paradox that S can reduce his loss in the general case $k > 2$ by accepting the obligation (21). The explanation lies in the circumstance that the

loss Λ has been computed as minimax, taking into account the possibility of D's playing his best method, which has been found to be the density $c(pq)^s$. In the general case he can not apply the same method for every face of the k-hedron because he would have to choose a density $f(p_1, p_2, \cdots, p_k)$ that would have the form $c_i p_i(1 - q_i)$ for $i = 1, 2, \cdots, k$ simultaneously, which is impossible. Thus the plurality of faces hinders the devil from simultaneously playing all the alternative games contained in the general case as well as possible. This obstacle for D opens the way for S to play every alternative game in the general case better than he could do it in the simple alternative $k = 2$; the inequality $4[(k - 1)/k] < k$ is the proof that the advantage thus gained by S is greater than the disadvantage that results from the condition (21). In fact we remarked just after (22) that this disadvantage is purely illusory.

9. The method of this paper applies as well to other distributions as to the binomial. The simplest example is given by the Poisson distribution. The problem is to estimate the mean number of signals per unit time, having observed an arbitrary unit interval and having found k signals in it. To apply the theory of games, we must define the amount to be paid for an erroneous estimate. If, to mention an example of analytical interest, this amount is equal to $(c - c')^2/c$, c being the true mean number and c' the estimate, the problem is easily solved.

Let us try the guess $c' = k$. The expected loss for a given c is

$$E = \sum_{k=0}^{\infty} \frac{(c - k)^2}{c} e^{-c} \cdot \frac{c^k}{k!} = cM_0 - 2M_1 + M_2/c, \tag{36}$$

the M_i $(i = 0, 1, 2)$ being the moments of order i of the Poisson distribution; this leads to

$$E = c - 2c + (\omega + c^2)/c = 1, \tag{37}$$

ω being the variance of the random variable k, which in this instance is c. Thus the loss E does not depend on c. We now have to find a prior distribution which leads to the method $c' = k$ as the relatively best. We will show that the uniform distribution does so. To speak correctly, it will be the distribution of constant density $T/(T^2 - 1)$ $(T > 1)$ in $(1/T, T)$ for $T \to \infty$. We have to compute x, for a given k so as to minimize the expected loss given by

$$\frac{T}{T^2 - 1} \int_{1/T}^{T} e^{-c} \cdot \frac{c^k(c - x)^2}{k!c} \, dc, \tag{38}$$

x being the guess we are seeking. The condition for the minimum is evidently

$$\int_{1/T}^{T} e^{-c} c^k \, dc = x \int_{1/T}^{T} e^{-c} c^{k-1} \, dc,$$

and becomes, for $T \to \infty$, $k! = x(k - 1)!$, $x = k$, Q.E.D.

It is interesting to note that no minimax solution exists if $(c - c')^2$ is fixed as the amount to be paid by S to D. The reason for this insolubility is to be sought in the freedom of the devil to increase c as far as he likes, putting for

instance $c_j = j$ or $c_j = 10^j$ in the jth round. It is easily seen that any guess c' far from k is worse for S than $c' = k$; on the other hand the expected value of $(c - k)^2$ is c, rising thus with c beyond every limit and increasing indefinitely the gain expected by D, whatever S does. The schedule of payments $(c - c')^2/c^2$ is also of practical interest, corresponding as it does to a penalty based on percentage error. Here again it is easy to see that no minimax solution exists.

10. Different objections to the new estimate have been raised:

1^0 The loss-function $(p - p')^2$ has been criticized as arbitrary. We have pointed out in the introduction that it is the result of an extrapolation of the estimate Ep of the old school; such a procedure is suggested by the circumstance that the symbol E loses its direct applicability when the prior distribution of p is either unknown or nonexistent.

2^0 The old estimate m/n has been compared with (10) as giving a better result, especially for n not too small. It has, however, been rejected not by mathematicians but by men of practice [7], who became aware of the paradox of error 0 for $m = 0$, $m = n$.

3^0 To save the minimax method, the Bayes hypothesis $f(p) \equiv 1$, and the estimate m/n, the loss-function $(p - p')^2/pq$ has been proposed, but no examples have been adduced for which this loss function seems realistic. There is also some interest in the loss function $(p - p')^2/(pq)^2$—for which no minimax estimate exists. There are examples in which this function corresponds to the real situation for p near 0 or for p near 1, but it is difficult to find natural problems in which the function measures the damage caused by bad guessing throughout the whole interval $0 < p < 1$.

4^0 The concept of a mathematically trained and satanically malicious devil has been denounced as leading to unnecessary caution; the estimate (10) is accused of paying too high a tribute to nature, which is incapable of seeking out such sophisticated devices against the mortals as given by (13). This objection may be answered as follows:

(a) We are compelled to play a game against nature ("the devil") and we do apply the new theory of games. If we are not allowed to do so we must abandon the new theory of games altogether and return to the old phraseology of "taking into account the psychology of the enemy." It is exactly the indefiniteness of such advice which made the new theory of games the only efficient tool in such problems as pursuit at sea; we need not worry whether the helmsman of the other boat is awake or not—he can just as well be an automaton and can be countered by our automaton.

(b) Let us drop the anthropomorphic idea of nature and admit that it acts blindly, without any strategy whatever. Such a view leads us naturally to deprive the devil of his sequence of coins of various frequencies. But even in such a situation it is difficult to imagine a better method for S against D than the new formulae: they do apply—not as minimax solutions but because of their independence of the p_j-sequence; they give a definite mean-square error, the only light in the total darkness.

5^0 Mr. A has been the first to investigate the blood-groups of a certain African tribe: he examined n' men and found m' with Rh-plus blood among them. Mr. B came second and examined another sample; his result was n'' and m'' respectively. Mr. B feels obliged to take into account the information contained in the paper published by A: he considers the result of his predecessor as furnishing a prior distribution and hesitates to apply (10) to his numbers. The sound procedure is, however, for B to publish his own numbers n'' and m'', to quote the numbers of A, and to estimate the Rh fraction of the tribe by setting $n = n' + n''$, $m = m' + m''$ in formula (10) with the remark: "making use of all available observations." This example illustrates the meaning of our arguments 4^0(a) and 4^0(b).

REFERENCES

[1] R. A. Fisher, *Statistical Methods and Scientific Inference*, Oliver and Boyd, Edinburgh, 1956.

[2] M. A. Girshick and D. Blackwell, *Theory of Games and Statistical Decisions*. John Wiley and Sons, New York, 1954.

[3] M. A. Girshick and L. J. Savage, "Bayes and minimax estimates for quadratic loss functions," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, 1951, pp. 53–74.

[4] J. L. Hodges, Jr., and E. L. Lehmann, "Some problems in minimax point estimation," *Ann. Math. Stat.*, Vol. 21 (1950), pp. 182–197.

[5] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, 2nd ed., Princeton University Press, Princeton, N. J., 1947, pp. 143–168.

[6] J. Neyman, "On the problem of confidence intervals," *Ann. Math. Stat.*, Vol. 6 (1935), pp. 111–116.

[7] A. M. Ritala, "Zur Berechnung des statistischen mittleren Fehlers (standard error)," *Acta Societatis Medicorum Fennicae "Duodecim,"* Ser. B, 19, fasc. 2, Helsinki, 1933.

[8] L. J. Savage, *The Foundations of Statistics*, John Wiley and Sons, New York, 1954, Sec. 13.4.

[9] H. Steinhaus, "Sur l'interprétation des résultats statistiques," *Colloquium Mathematicum*, Vol. 1 (1948), pp. 232–238.

[10] H. Steinhaus, "Uber einige prinzipielle Fragen der mathematischen Statistik," *Tagung über Wahrscheinlichkeitsrechnung und mathematische Statistik, Berlin 19–21. X. 1954*, Deutscher Verlag der Wissenschaften, V: Das statistische Spiel.

[11] S. Trybuła, "Some problems of simultaneous minimax estimation," submitted for publication in *Ann. Math. Stat.*

[12] A. Wald, *Statistical Decision Functions*, John Wiley and Sons, New York, 1950.

[13] E. T. Whittaker and G. N. Watson, *Modern Analysis*, 4th ed., Cambridge University Press, Cambridge, 1940, Sec. 12.5.

NON-PARAMETRIC EMPIRICAL BAYES PROCEDURES[1]

BY M. V. JOHNS, JR.

Columbia University

1. Introduction and summary. In the usual formulation of problems in statistical decision theory the probability distribution of the observations is assumed to be a member of some specified class of distribution functions. No a priori information is ordinarily assumed to exist concerning which member of this class is the true distribution of the observations although a priori probability measures defined over this class may be introduced as a technical device for generating complete classes of decision functions, minimax decision rules, etc. However, in some experimental situations it may be reasonable to suppose that such an a priori probability measure actually exists in the sense that the distributions of observations occurring in different experiments made under similar circumstances may be thought of as having been selected from a specified class of distribution functions according to some probability law.

Such an assumption seems particularly apt in the case where measurements are made on an individual selected according to some probability law (e.g., "at random") from a population and where it is desired to make inferences about some characteristic of the individual on the basis of these measurements. If the class of probability distributions of the measurements for all individuals in the population and the law of selection are known, an optimum Bayes decision procedure can then be found. In general, however, such information will not be available to the experimenter, but there may be observations available on individuals previously selected in the same way from the same population and, under certain circumstances, these prior observations may be used to obtain approximations to the optimum Bayes decision procedure. The possibility of using prior observations to approximate Bayes procedures was first established for certain estimation problems in [1] by H. Robbins who coined the term "empirical Bayes procedures" to describe such approximations.

Robbins in [1] discusses the estimation, using a squared error loss function, of the value λ of a random variable Λ associated with a discrete valued observation X whose conditional probability function, given λ, is $p(x \mid \lambda)$, where $p(x \mid \lambda)$ is known for each λ but where the (a priori) distribution of Λ is unknown. For several specific parametric families of discrete probability functions $p(x \mid \lambda)$ Robbins shows that if prior independent observations $X_1, X_2, \cdots, X_n$, each having the same unconditional distribution as X are available, then an empirical Bayes estimator using $X_1, X_2, \cdots, X_n$ can be found which converges with probability 1 to the Bayes estimator as n increases, for any a priori distribution of Λ.

Received July 2, 1956.

[1] Work sponsored by the Office of Naval Research, Nonr Contract No. 266(33).

In Sec. 2 below a similar estimation problem is considered for the "nonparametric" case where the class of (conditional) probability distributions of X is not restricted to a particular parametric family, but is instead the class of all probability functions assigning probability 1 to some specified denumerable set of numbers. The quantity to be estimated is the value of a functional defined on this class of probability functions and it is assumed that there exists an unknown a priori probability measure defined on a suitable σ-algebra of subsets of this class. For this case it is shown that under certain circumstances prior observations may be used to construct empirical Bayes estimators having the property that, as the number of prior observations increases, the risks of the empirical Bayes estimators converge to the risk of the Bayes estimator for any a priori probability measure provided that certain moments exist. The rate of convergence of these risks is also investigated for two special cases.

In Sec. 3 the techniques of Sec. 2 are modified to apply to the case where the class of (conditional) distributions of X is the class of all absolutely continuous distribution functions, and similar results are obtained.

In Sec. 4 the results of the previous sections are used to obtain empirical Bayes solutions for certain two-decision problems of the hypothesis-testing type.

Throughout this paper certain elementary properties of conditional expectations are used which are immediate consequences of results contained, for example, in Chapter VII of [2].

2. Estimation: the discrete case. For a specified denumerable set of numbers $\chi = \{x\}$ let $\mathfrak{F} = \{F(x \mid \omega) : \omega \in \Omega\}$ be the class of all c.d.f.'s assigning probability 1 to χ, where $\Omega = \{\omega\}$ is an abstract indexing set. Let μ be an a priori probability measure defined on a σ-algebra $\mathfrak{A}$ of subsets of Ω, and let Y be the Ω-valued random variable which is the identity mapping of Ω onto itself. We may then define the random variables $X_1, X_2, \cdots, X_r$ so that they are conditionally independently and identically distributed with the common c.d.f. $F(x \mid \omega)$ given that $Y = \omega$. Finally, for a given measurable function $h(x)$ we define the random variable Λ by

$$\Lambda = \Lambda(Y) = E(h(X) \mid Y), \tag{2.1}$$

where X is a generic representative of the X_j's. We assume that the a priori probability measure space $(\Omega, \mathfrak{A}, \mu)$ is such that

$$Eh^2(X) < \infty. \tag{2.2}$$

Here Λ may be thought of as a functional defined on $\mathfrak{F}$ or, equivalently, as a function defined on Ω. We might, for example, define $h(x) \equiv x$ so that the value of $\Lambda(\omega)$ is the expected value of X given that the c.d.f. of X is $F(x \mid \omega)$. Another possibility would be to let $h(x) = 1$ if $x < c$, and $h(x) = 0$ otherwise, so that $\Lambda(\omega) = F(c \mid \omega)$.

Suppose that we wish to use the vector of observations

$$\underset{\sim}{X} = (X_1, X_2, \cdots, X_r)$$

to obtain an estimate of the value λ assumed by Λ, where the loss incurred when t is the estimated value is

$$L(t, \lambda) = (t - \lambda)^2. \tag{2.3}$$

Then the risk involved in using any estimator $\varphi(\underset{\sim}{x})$ is

$$R(\varphi) = EL(\varphi(\underset{\sim}{X}), \Lambda) = E[\varphi(\underset{\sim}{X}) - \Lambda]^2. \tag{2.4}$$

Now from (2.1) amd (2.2) we have

$$E\Lambda^2 = E\{E^2(h(X) \mid Y)\} \leqq Eh^2(X) < \infty, \tag{2.5}$$

so that $R(\varphi)$ may be written

$$\begin{aligned} &R(\varphi) \\ &= E\{[\varphi^2(X) - 2\varphi(X)E(\Lambda \mid \underset{\sim}{X}) + E^2(\Lambda \mid \underset{\sim}{X})] + E(\Lambda^2 \mid \underset{\sim}{X}) - E^2(\Lambda \mid \underset{\sim}{X})\}. \end{aligned} \tag{2.6}$$

The expression in square brackets is a perfect square and hence is non-negative and equals zero if and only if $\varphi(X) = E(\Lambda \mid \underset{\sim}{X})$. Thus the Bayes estimator $\varphi_B(\underset{\sim}{x})$ which minimizes $R(\varphi)$ is given by

$$\varphi_B(\underset{\sim}{x}) = E(\Lambda \mid \underset{\sim}{X} = \underset{\sim}{x}) \tag{2.7}$$

for all $\underset{\sim}{x} = (x_1, x_2, \cdots, x_r) \varepsilon \chi^* = \{\underset{\sim}{x}: \text{Prob}\,(\underset{\sim}{X} = \underset{\sim}{x}) > 0\}$. The risk of the Bayes estimator is then

$$R(\varphi_B) = E\Lambda^2 - E\varphi_B^2(X) < \infty. \tag{2.8}$$

To obtain the Bayes estimator φ_B we must, of course, know the a priori probability structure of the problem. Suppose now that this structure is *unknown* but that collateral information is available, the form of which is determined as follows:

Let $Y_1, Y_2, \cdots, Y_n$ be mutually independent Ω-valued random variables each of which is independent of Y and has the same distribution as Y. Then let the additional information be in the form of vectors of observations $\underset{\sim}{X}_i = (X_{i1}, X_{i2}, \cdots, X_{i,r+1})$, $i = 1, 2, \cdots, n$ where the $\underset{\sim}{X}_i$'s are mutually independent and independent of $\underset{\sim}{X}$ and where for each i the X_{ij}'s are conditionally independent and identically distributed according to $F(x \mid \omega_i)$ given that $Y_i = \omega_i$. Here although the $\underset{\sim}{X}_i$'s are independent of $\underset{\sim}{X}$ and Λ, they nevertheless contain useful information since they possess the same a priori probability structure as $\underset{\sim}{X}$. Thus, if we let $\underset{\sim}{X}_i^{(r)} = (X_{i1}, X_{i2}, \cdots, X_{ir})$ then $\underset{\sim}{X}_i^{(r)}$ and

$$E(h(X_{i,r+1}) \mid Y_i)$$

have the same joint distribution as X and Λ so that

$$\begin{aligned} E(h(X_{i,r+1}) \mid X_i^{(r)} = x) &= E\{E(h(X_{i,r+1}) \mid Y_i, X_i^{(r)}) \mid X_i^{(r)} = x\} \\ &= E\{E(h(X_{i,r+1}) \mid Y_i) \mid X_i^{(r)} = x\} \\ &= E(\Lambda \mid X = x) = \varphi_\mu(x), \end{aligned} \tag{2.9}$$

for $x \in \chi^*$. This suggests the following empirical Bayes estimation procedure:

Let $x_{(q)}$, $q = 1, 2, \cdots, m(x)$ represent the $m(x)$ distinct vectors obtained by permuting the components $x_1, x_2, \cdots, x_r$ of x. Clearly for each x we have $1 \leqq m(x) \leqq r!$. Now we define the random functions $M_i(x)$, $i = 1, 2, \cdots, n$, and $\bar{M}_n(x)$ by

$$M_i(x) = \begin{cases} 1, & \text{if there exists a } q,\ 1 \leqq q \leqq m(x), \\ & \text{such that } X_i^{(r)} = x_{(q)}, \\ 0, & \text{otherwise,} \end{cases} \tag{2.10}$$

and

$$\bar{M}_n(x) = \sum_{i=1}^{n} M_i(x). \tag{2.11}$$

Then we define an empirical Bayes estimator $\varphi_n(x)$ by

$$\varphi_n(x) = \begin{cases} \dfrac{1}{\bar{M}_n(x)} \displaystyle\sum_{i=1}^{n} M_i(x) h(X_{i,r+1}), & \bar{M}_n(x) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2.12}$$

In order to show that the risk involved in using φ_n approaches the risk for the Bayes estimator φ_μ as n becomes large we first prove two lemmas. For $n = 1, 2, \cdots$, let

$$P_n(x) = \text{Prob}\,\{\bar{M}_n > 0\}, \tag{2.13}$$

$$V_n(x) = \begin{cases} \dfrac{1}{\bar{M}_n(x)}, & \bar{M}_n(x) > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{2.14}$$

$$\xi_n(x) = EV_n(x), \tag{2.15}$$

and for $x \in \chi^*$,

$$\theta(x) = E(h^2(X_{i,r+1}) \mid X_1^{(r)} = x). \tag{2.16}$$

LEMMA 1. *For any fixed $x \in \chi^*$,*

$$E\varphi_n(x) = \varphi_\mu(x) P_n(x), \tag{2.17}$$

and

$$E\varphi_n^2(x) = [\theta(x) - \varphi_\mu^2(x)]\xi_n(x) + \varphi_\mu^2(x) P_n(x). \tag{2.18}$$

Proof. For notational simplicity we let $H_i = h(X_{i,r+1})$ and suppress the fixed argument x ($\varepsilon \chi^*$) in all functions defined above. Then letting

$$\underline{U}_n = (M_1, M_2, \cdots, M_n)$$

and noting that for each i the joint distribution of the X_{ij}'s is invariant under permutations we have

$$E(M_iH_i \mid \underline{U}_n) = E(M_iH_i \mid M_i) = M_i\varphi_\mu, \tag{2.19}$$

by (2.9), and

$$E(M_i^2H_i^2 \mid \underline{U}_n) = E(M_i^2H_i^2 \mid M_i) = M_i\theta. \tag{2.20}$$

Also for $i \neq q$,

$$\begin{aligned} E(M_iM_qH_iH_q \mid \underline{U}_n) &= E\{M_iH_iE(M_qH_q \mid H_i, M_i, M_q) \mid M_i, M_q\} \\ &= E\{M_iH_iE(M_qH_q \mid M_q) \mid M_i, M_q\} \\ &= E(M_iH_i \mid M_i)E(M_qH_q \mid M_q) = M_iM_q\varphi_\mu^2. \end{aligned} \tag{2.21}$$

Hence

$$E\varphi_n = E\left\{\frac{1}{\bar{M}_n}\sum_{i=1}^{n} E(M_iH_i \mid \underline{U}_n) \,\middle|\, \bar{M}_n > 0\right\} P_n = \varphi_\mu P_n, \tag{2.22}$$

and

$$\begin{aligned} E\varphi_n^2 &= E\left\{\frac{1}{\bar{M}_n^2}\sum_{i=1}^{n} E(M_i^2H_i^2 \mid \underline{U}_n) \,\middle|\, \bar{M}_n > 0\right\} P_n \\ &\quad + E\left\{\frac{1}{\bar{M}_n^2}\sum_{i\neq q}^{n}\sum^{n} E(M_iM_qH_iH_q \mid \underline{U}_n) \,\middle|\, \bar{M}_n > 0\right\} P_n, \end{aligned} \tag{2.23}$$

$$= \theta E\left\{\frac{1}{\bar{M}_n} \,\middle|\, \bar{M}_n > 0\right\} P_n + \varphi_\mu^2 E\left\{\frac{1}{\bar{M}_n^2}\sum_{i\neq q}^{n}\sum^{n} M_iM_q \,\middle|\, \bar{M}_n > 0\right\} P_n.$$

Now since

$$\frac{1}{\bar{M}_n^2}\sum_{i\neq q}^{n}\sum^{n} M_iM_q = 1 - \frac{1}{\bar{M}_n}, \tag{2.24}$$

and

$$\xi_n = EV_n = E\left\{\frac{1}{\bar{M}_n} \,\middle|\, \bar{M}_n > 0\right\} P_n, \tag{2.25}$$

we have

$$E\varphi_n^2 = \theta\xi_n + \varphi_\mu^2[P_n - \xi_n], \tag{2.26}$$

and the proof of the lemma is complete.

LEMMA 2. *For any $x \varepsilon \chi^*$,*

$$\lim_{n\to\infty} P_n(x) = 1, \tag{2.27}$$

and

$$\lim_{n\to\infty} \xi_n(\underline{x}) = 0. \tag{2.28}$$

Proof. Let

$$p(\underline{x}) = \text{Prob}\,\{M_1(\underline{x}) = 1\}. \tag{2.29}$$

Then $\underline{x} \in \chi^*$ implies $p(\underline{x}) > 0$. Now for any fixed $\underline{x}$, $\bar{M}_n(\underline{x})$ is a binomial variable with parameters n and $p(\underline{x})$. Hence for sufficiently large n, $\bar{M}_n(\underline{x})$ will be arbitrarily large with probability arbitrarily close to 1 for any $\underline{x} \in \chi^*$. This implies that for $\underline{x} \in \chi^*$,

$$\lim_{n\to\infty} P_n(\underline{x}) = \lim_{n\to\infty} \text{Prob}\,\{\bar{M}_n(\underline{x}) > 0\} = 1, \tag{2.30}$$

and

$$V_n(\underline{x}) \to 0, \qquad \text{in probability,} \tag{2.31}$$

as $n \to \infty$. Now (2.31), together with the fact that $0 \leqq V_n(\underline{x}) \leqq 1$, implies

$$\lim_{n\to\infty} \xi_n(\underline{x}) = \lim_{n\to\infty} EV_n(\underline{x}) = 0, \tag{2.32}$$

for $\underline{x} \in \chi^*$, which completes the proof of the lemma.

We now prove the following theorem:

THEOREM 1. *If the a priori probability measure space* $(\Omega, \mathcal{A}, \mu)$ *is such that* (2.2) *is satisfied, then*

$$\lim_{n\to\infty} R(\varphi_n) = R(\varphi_B). \tag{2.33}$$

Proof. We first observe that

$$R(\varphi_n) = E[\varphi_n(\underline{X}) - \Lambda]^2 = E\varphi_n^2(\underline{X}) - 2E[\Lambda\varphi_n(\underline{X})] + E\Lambda^2, \tag{2.34}$$

provided that all of the terms on the right are finite. Now since $\underline{X}$ and Λ are independent of $X_1, X_2, \cdots, X_n$, we have

$$E(\varphi_n^2(\underline{X}) \mid \underline{X} = \underline{x}) = E\varphi_n^2(\underline{x}), \tag{2.35}$$

and

$$\begin{aligned} E(\Lambda\varphi_n(\underline{X}) \mid \underline{X} = \underline{x}) &= E\{\Lambda E(\varphi_n(\underline{X}) \mid \Lambda, \underline{X}) \mid \underline{X} = \underline{x}\} \\ &= E\{\Lambda E(\varphi_n(\underline{X}) \mid \underline{X}) \mid \underline{X} = \underline{x}\} \\ &= \varphi_B(\underline{x})E\varphi_n(\underline{x}), \end{aligned} \tag{2.36}$$

for all $\underline{x} \in \chi^*$. Hence by (2.17) and (2.18) of Lemma 1 together with (2.34), (2.35) and (2.36) we have

$$R(\varphi_n) = E\Lambda^2 + E[\theta(\underline{X})\xi_n(\underline{X})] - E[\varphi_B^2(\underline{X})(P_n(\underline{X}) + \xi_n(\underline{X}))]. \tag{2.37}$$

Now by (2.27) and (2.28) of Lemma 2, for all $\underline{x} \in \chi^*$,

$$\lim_{n\to\infty} \theta(x)\xi_n(x) = 0, \tag{2.38}$$

and

$$\lim_{n\to\infty} \varphi_n^2(x)(P_n(x) + \xi_n(x)) = \varphi_\mu^2(x). \tag{2.39}$$

Also since $0 \leqq \xi_n(x) \leqq 1$ and $0 \leqq P_n(x) \leqq 1$, we have

$$|\theta(x)\xi_n(x)| \leqq \theta(x), \tag{2.40}$$

and

$$|\varphi_n^2(x)(P_n(x) + \xi_n(x))| \leqq 2\varphi_\mu^2(x), \tag{2.41}$$

and furthermore since (2.2) is satisfied,

$$E\theta(X) = Eh^2(X) < \infty, \tag{2.42}$$

and

$$E\varphi_\mu^2(X) < \infty. \tag{2.43}$$

Hence by the Lebesgue Dominated Convergence Theorem we may assert

$$\lim_{n\to\infty} R(\varphi_n) = E\Lambda^2 - E\varphi_\mu^2(X) = R(\varphi_\mu), \tag{2.44}$$

which is the desired result.

This result is "non-parametric" in the sense that we have assumed that the unknown a priori probability measure is defined over the class $\mathfrak{F}$ of *all* c.d.f.'s assigning probability 1 to the set χ. If we are willing to assume that some specific parametric subclass of $\mathfrak{F}$ is assigned *a priori* probability 1 then we may be able to find empirical Bayes estimators such as those discussed by Robbins in [1] which (presumably) are more efficient than (2.12) for such cases.

It seems likely that the empirical Bayes estimator φ_n given by (2.12) is relatively inefficient when n is small or r is large relative to n, since, in this case, $\bar{M}_n(X)$ is small with high probability so that relatively few of the X_i's contribute useful information to φ_n. This difficulty may be offset to some extent by replacing φ_n by an estimate of Λ based on the value of X when $\bar{M}_n(X)$ is small. Thus, for example, we may define a modified empirical Bayes estimator $\varphi_n^{(c)}$ for fixed $c \geqq 0$ by

$$\varphi_n^{(c)}(x) = \begin{cases} \varphi_n(x), & \bar{M}_n(x) > c, \\ \dfrac{1}{r}\displaystyle\sum_{j=1}^{r} h(x_j), & \bar{M}_n(x) \leqq c. \end{cases} \tag{2.45}$$

It is not difficult to verify that $\varphi_n^{(c)}$ has the same asymptotic properties as φ_n and that for small n, $R(\varphi_n^{(c)})$ tends to be smaller than $R(\varphi_n)$ except when the distribution of Λ is concentrated near zero.

If the vectors of prior observations X_i are of the form $X_i = (X_{i1}, X_{i2}, \cdots,$

$X_{i,r+k_i}$) where $k_i \geqq 1$, $i = 1, 2, \cdots, n$, we may make use of the additional information available when $k_i > 1$ by substituting $M_i(\underline{x})(w_i/k_i) \sum_{j=r+1}^{k_i} h(X_{i,j})$ for $M_i(\underline{x})h(X_{i,r+1})$ and $\sum_{i=1}^{n} w_i M_i(\underline{x})$ for $\bar{M}_n(\underline{x})$ in (2.11), where $w_1, w_2, \cdots, w_n$ are positive numerical weights depending on the k_i's. It can be shown by arguments similar to those of Theorem 1 that whenever the w_i's are uniformly bounded above and below by positive constants then the risk of the resulting estimator approaches $R(\varphi_\mu)$ as n becomes large. However, in general when the k_i's are not all the same the optimum choice of $w_1, w_2, \cdots, w_n$ depends on the unknown a priori probability structure and hence cannot be determined.

In practice it may happen that the numbers of components in the $\underline{X}_i$'s and in $\underline{X}$ are all the same ($= r + 1$, say) so that in order to use the estimator φ_n given by (2.12) we must discard one of the components of $\underline{X}$. This seems undesirable and suggests the use of the modified empirical Bayes estimator

$$\tilde{\varphi}_n(\underline{x}) = \frac{1}{r+1} \sum_{j=1}^{r+1} \varphi_n(\underline{x}^{(j)}), \tag{2.46}$$

where $\underline{x}^{(j)} = (x_1, x_2, \cdots, x_{j-1}, x_{j+1}, \cdots, x_{r+1})$, $j = 1, 2, \cdots, r+1$ and φ_n is given by (2.12). To evaluate the performance of $\tilde{\varphi}_n(\underline{x})$ we compare its risk with that of $\varphi_n(\underline{x}^{(1)})$ which uses only r components of $\underline{X}$, as follows: For all n,

$$\begin{aligned} R(\tilde{\varphi}_n) &= E\Lambda^2 + E\tilde{\varphi}_n^2(\underline{X}) - 2E[\Lambda\tilde{\varphi}_n(\underline{X})] \\ &= E\Lambda^2 + \frac{1}{r+1} E\varphi_n^2(\underline{X}^{(1)}) \\ &\qquad + \frac{r}{r+1} E[\varphi_n(\underline{X}^{(1)})\varphi_n(\underline{X}^{(2)})] - 2E[\Lambda\varphi_n(\underline{X}^{(1)})] \\ &= R(\varphi_n) - \frac{1}{2}\frac{r}{r+1} E[\varphi_n(\underline{X}^{(1)}) - \varphi_n(\underline{X}^{(2)})]^2 \\ &\leqq R(\varphi_n), \end{aligned} \tag{2.47}$$

with equality holding only in the degenerate case where $\underline{X}^{(1)} = \underline{X}^{(2)}$ with probability 1. By arguments similar to those of Theorem 1 it may be shown that

$$\lim_{n\to\infty} R(\tilde{\varphi}_n) = R(\tilde{\varphi}_\mu) \leqq R(\varphi_\mu), \tag{2.48}$$

where $R(\tilde{\varphi}_\mu)$ is the risk of the estimator

$$\tilde{\varphi}_\mu(\underline{x}) = \frac{1}{r+1} \sum_{j=1}^{r+1} E(\Lambda \mid \underline{X}^{(j)} = \underline{x}^{(j)}), \tag{2.49}$$

and $R(\varphi_\mu)$ is the risk of the ordinary Bayes estimator $\varphi_\mu(\underline{x}^{(1)}) = E(\Lambda \mid \underline{X}^{(1)} = \underline{x}^{(1)})$ based on only r of the $r + 1$ components of $\underline{X}$. The inequality in (2.48) will again be strict except in degenerate cases.

It would be interesting to compute the value of $R(\varphi_n)$ for various values of n for some specific a priori probability distribution in order to obtain some idea

of the rate at which $R(\varphi_n)$ converges to $R(\varphi_\mu)$. Unfortunately, such computations are extremely lengthy even in the simplest non-trivial cases. However, reasonably good upper and lower bounds for $R(\varphi_n)$ may be computed without much difficulty by using simple bounds for $\xi_n(\underline{x})$ obtained as follows:

$$\xi_n(\underline{x}) = EV_n(\underline{x}) = \sum_{s=1}^{n} \frac{1}{s}\, b(s; n, p(\underline{x})) \tag{2.50}$$

where $b(s; n, p(\underline{x}))$ is the probability of s successes occurring in n independent trials where the probability of success at each trial is $p(\underline{x}) = \text{Prob}\,\{M_1(\underline{x}) = 1\}$. Also

$$\begin{aligned}\sum_{s=1}^{n} \frac{1}{s+1}\, b(s; n, p(\underline{x})) &= \frac{1}{(n+1)p(\underline{x})} \sum_{j=1}^{n} b(s+1; n+1, p(\underline{x})) \\ &= \frac{1}{(n+1)p(\underline{x})} \{1 - [1 + np(\underline{x})][1 - p(\underline{x})]^n\},\end{aligned} \tag{2.51}$$

for $\underline{x} \in \chi^*$. Hence, letting

$$b_n(\underline{x}) = \frac{1}{(n+1)p(\underline{x})} \{1 - [1 + np(\underline{x})][1 - p(\underline{x})]^n\}, \tag{2.52}$$

and noting that $1/(s+1) < 1/s \leqq 2/(s+1) < 1$ for all $s \geqq 1$, we have

$$b_n(\underline{x}) < \xi_n(\underline{x}) \leqq 2b_n(\underline{x}) \leqq 1, \tag{2.53}$$

for $\underline{x} \in \chi^*$.

Suppose now that we wish to estimate the expected value of a non-negative integer-valued random variable X (i.e., we set $h(x) \equiv x$ and $\chi = \{0, 1, 2, \cdots\}$), and suppose that $r = 1$ so that $\underline{X} = X$ and $\underline{X}_i = (X_{i1}, X_{i2})\ i = 1, 2, \cdots, n$. For this problem we may compute the upper bounds based on (2.53) for the risks of the estimators φ_n and $\varphi_n^{(0)}$ given respectively by (2.12) and (2.45) with $c = 0$, for the particular a priori probability measure μ which assigns probability 1 to the family of Poisson c.d.f.'s

$$F(x \mid \lambda) = \sum_{t=0}^{[x]} \frac{1}{t!} e^{-\lambda t} \lambda^t; \qquad x \geqq 0, \qquad \lambda > 0, \tag{2.54}$$

and which induces the Γ-distribution

$$G(\lambda) = \int_0^\lambda \frac{\alpha^\beta}{\Gamma(\beta)}\, u^{\beta-1} e^{-\alpha u}\, du; \qquad \alpha, \beta > 0, \tag{2.55}$$

as the c.d.f. of Λ. In Table I below we compare the upper and lower bounds for $R(\varphi_n)$ and $R(\varphi_n^{(0)})$ with the risk $R(\varphi_\mu)$ of the Bayes estimator φ_μ and with the risk $R(x)$ of the classical estimator x. These quantities are computed for six values of n, for $\alpha = 2$ and $\beta = 10$ (i.e., $E\Lambda = 5$ and $\text{Var}\Lambda = 2.5$). In Table II we compare the same quantities for the case where μ assigns a priori probability 1 to the particular member of the family of c.d.f.'s (2.54) for which $\lambda = \lambda_0 = 5$ (i.e., $\Lambda \equiv \lambda_0$ with probability 1).

TABLE I

n	$R(\varphi_n)$		$R(\varphi_n^{(2)})$		$R(\varphi_n)$	$R(x)$
	Lower Bound	Upper Bound	Lower Bound	Upper Bound		
15	10.70	12.51	5.21	7.02	1.67	5.00
30	6.74	8.32	4.43	6.01	1.67	5.00
60	4.46	5.54	3.50	4.58	1.67	5.00
120	3.20	3.87	2.79	3.46	1.67	5.00
240	2.50	2.89	2.32	2.72	1.67	5.00
480	2.12	2.34	2.05	2.27	1.67	5.00

TABLE II

n	$R(\varphi_n)$		$R(\varphi_n^{(2)})$		$R(\varphi_n)$	$R(x)$
	Lower Bound	Upper Bound	Lower Bound	Upper Bound		
15	6.06	7.44	3.68	5.06	0	5.00
30	2.94	4.04	2.57	3.66	0	5.00
60	1.46	2.17	1.57	2.28	0	5.00
120	0.74	1.16	0.91	1.32	0	5.00
240	0.38	0.61	0.51	0.74	0	5.00
480	0.19	0.32	0.27	0.40	0	5.00

3. Estimation: the continuous case. In this section we extend the methods of Sec. 2 to cover the case where the observed X's possess absolutely continuous distribution functions. As before these results will be "non-parametric" in the sense that the unknown a priori probability measure is assumed to be defined over the class of all absolutely continuous c.d.f.'s subject only to some conditions on the existence of moments.

Let $\mathfrak{F} = \{F(x \mid \omega) : \omega \in \Omega\}$ be the collection of all absolutely continuous c.d.f.'s where $\Omega = \{\omega\}$ is an abstract indexing set, and let $(\Omega, \mathfrak{A}, \mu)$ be an a priori probability measure space where $\mathfrak{A}$ is a σ-algebra of subsets of Ω. Then there exists a function $f(u \mid \omega)$ defined on (reals) $\times\ \Omega$ such that

$$F(x \mid \omega) = \int_{-\infty}^{x} f(u \mid \omega)\, du, \tag{3.1}$$

for each $\omega \in \Omega$. We assume that $(\Omega, \mathfrak{A})$ is such that the function $f(u \mid \omega)$ is a measurable function on the product space (reals) $\times\ \Omega$.

Let $Y = Y(\omega)$ be the Ω-valued random variable which is the identity mapping of Ω onto itself. Let the (real) random variables $X_1, X_2, \cdots, X_r$ be conditionally independent and identically distributed according to $F(x \mid \omega)$ given that $Y = \omega$, and let $\underline{X} = (X_1, X_2, \cdots, X_r)$. Then the unconditional joint c.d.f. of $X_1, X_2, \cdots, X_r$ will be

$$F(x_1, \cdots, x_r) = F(\underline{x}) = \int_{\Omega} \prod_{j=1}^{r} F(x_j \mid \omega)\, d\mu$$

$$= \int_\Omega \prod_{j=1}^{r} \int_{-\infty}^{x_j} f(u_j \mid \omega)\, du_j\, d\mu \tag{3.2}$$

$$= \int_{-\infty}^{x_r} \cdots \int_{-\infty}^{x_1} f(u_1, \cdots, u_r)\, du_1 \cdots du_r,$$

where

$$f(x_1, \cdots, x_r) = f(\underline{x}) = \int_\Omega \prod_{j=1}^{r} f(x_j \mid \omega)\, d\mu. \tag{3.3}$$

Thus $f(\underline{x})$ is the joint unconditional probability density function of $X_1, \cdots, X_r$. Now for a given measurable function $h(x)$ we define the random variable Λ by

$$\Lambda = \Lambda(Y) = E(h(X) \mid Y), \tag{3.4}$$

where X is a generic representative of the X_j's. As in Sec. 2 we assume that

$$Eh^2(X) < \infty, \tag{3.5}$$

which implies $E\Lambda^2 < \infty$ and hence the existence of all conditional expectations with which we will be concerned. Now if we wish to estimate the value of Λ using $\underline{X}$ where the risk is the expected squared error, then, as before, the essentially unique Bayes estimator is

$$\psi_n(\underline{x}) = E(\Lambda \mid \underline{X} = \underline{x}). \tag{3.6}$$

(In this section when conditional expectations are regarded as functions of the values of random variables it will be understood that we mean the essentially unique Borel measurable version which is set equal to zero whenever arbitrariness is possible on sets of positive Lebesgue measure.)

As in Sec. 2 we introduce random variables $Y_1, \cdots, Y_n$ independent of each other and of Y such that each has the same distribution as Y. We also introduce the random vectors of prior observations $\underline{X}_i = (X_{i1}, \cdots, X_{i,r+1})$, $i = 1, 2, \cdots, n$, where the $\underline{X}_i$'s are independent of each other and of $\underline{X}$ and where for each i the X_{ij}'s are conditionally independent and identically distributed according to $F(x \mid \omega_i)$ given that $Y_i = \omega_i$. As before we let $\underline{X}_i^{(r)} = (X_{i1}, X_{i2}, \cdots, X_{ir})$, $i = 1, 2, \cdots, n$, and we note that for each i

$$E(h(X_{i,r+1}) \mid \underline{X}_i^{(r)} = \underline{x}) = E(\Lambda \mid \underline{X} = \underline{x}) = \psi_n(\underline{x}). \tag{3.7}$$

In order to make use of the results of Sec. 2 we must discretize the X's in some way. To this end we consider the double sequence of half-open intervals

$$I_t^{(n)} = \left[\frac{tc}{n^{1-\delta/r}}, \frac{(t+1)c}{n^{1-\delta/r}}\right), \qquad t = 0, \pm 1, \pm 2, \cdots; \qquad n = 1, 2, \cdots, \tag{3.8}$$

where $c > 0$ and $0 < \delta < 1$. For each n we partition r-dimensional euclidean space into a countable sequence of non-overlapping hypercubes $C_j^{(n)}$, $j = 1, 2, \cdots$, of the form

$$C_j^{(n)} = I_{t_{1j}}^{(n)} \times I_{t_{2j}}^{(n)} \times \cdots \times I_{t_{rj}}^{(n)}, \qquad j = 1, 2, \cdots, \tag{3.9}$$

where the t_{ij}'s are suitably chosen integers. Then for each n and each r-component numerical vector $\underline{x} = (x_1, x_2, \cdots, x_r)$ we let $C^{(n)}(\underline{x})$ be the unique member of the sequence (3.9) containing $\underline{x}$. As before, we designate by $\underline{x}_{(q)}$, $q = 1, 2, \cdots, m(\underline{x})$, $m(\underline{x}) \geqq 1$, the distinct vectors obtained by permuting the components of $\underline{x}$. Then proceeding by analogy with Sec. 2 we define the random functions

$$M_i^{(n)}(\underline{x}) = \begin{cases} 1, & \text{if there exists a } q,\ 1 \leqq q \leqq m(\underline{x}), \\ & \text{such that } X_i^{(r)} \varepsilon\, C^{(n)}(\underline{x}_{(q)}), \\ 0, & \text{otherwise,} \end{cases} \tag{3.10}$$

for $i = 1, 2, \cdots, n$, and

$$\bar{M}^{(n)}(\underline{x}) = \sum_{i=1}^{r} M_i^{(n)}(\underline{x}), \tag{3.11}$$

and finally we define the empirical Bayes estimator $\psi_n(\underline{x})$ by

$$\psi_n(\underline{x}) = \begin{cases} \dfrac{1}{\bar{M}^{(n)}(\underline{x})} \displaystyle\sum_{i=1}^{n} M_i^{(n)}(\underline{x}) h(X_{i,r+1}), & \bar{M}^{(n)}(\underline{x}) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.12}$$

Before we can show that $\lim_{n\to\infty} R(\psi_n) = R(\psi_\mu)$ we must prove three preliminary lemmas.

We first let

$$\varphi^{(n)}(\underline{x}) = E(h(X_{1,r+1}) \mid M_1^{(n)}(\underline{x}) = 1), \qquad n = 1, 2, \cdots, \tag{3.13}$$

and

$$\theta^{(n)}(\underline{x}) = E(h^2(X_{1,r+1}) \mid M_1^{(n)}(\underline{x}) = 1), \qquad n = 1, 2, \cdots. \tag{3.14}$$

For each n, $\varphi^{(n)}(\underline{x})$ and $\theta^{(n)}(\underline{x})$ are analogous respectively to $\varphi_\mu(\underline{x})$ and $\theta(\underline{x})$ of Sec. 2. We now prove the following lemma:

LEMMA 3. *For almost all $\underline{x}$ such that $f(\underline{x}) > 0$,*

$$\lim_{n\to\infty} \varphi^{(n)}(\underline{x}) = E(h(X_{1,r+1}) \mid X_1^{(r)} = \underline{x}) = \psi_\mu(\underline{x}), \tag{3.15}$$

and

$$\lim_{n\to\infty} \theta^{(n)}(\underline{x}) = E(h^2(X_{1,r+1}) \mid X_1^{(r)} = \underline{x}). \tag{3.16}$$

Proof. Let the joint density function for the $r + 1$ components of X_i (for any fixed i) be written as

$$f(\underline{x}, x_{r+1}) = f(x_1, x_2, \cdots, x_{r+1}) = \int_\Omega \prod_{j=1}^{r+1} f(x_j \mid \omega)\, d\mu. \tag{3.17}$$

Then since $f(\underline{x}, x_{r+1}) = f(\underline{x}_{(q)}, x_{r+1})$, $q = 1, 2, \cdots, m(\underline{x})$, we have for almost

all $\underline{x}$

$$\varphi^{(n)}(\underline{x}) = \frac{\int_{C^{(n)}(\underline{x})} \int_{-\infty}^{\infty} h(v) f(\underline{y}, v)\, dv\, d\underline{y}}{\int_{C^{(n)}(\underline{x})} f(\underline{y})\, d\underline{y}}, \tag{3.18}$$

provided the denominator is not zero. (If the denominator is zero, we conventionally set $\varphi^{(n)}(\underline{x}) = 0$). Now for each $\underline{x}$, $C^{(n)}(\underline{x})$, $n = 1, 2, \cdots$, is a sequence of hypercubes converging regularly to the point $\underline{x}$. Hence for almost all $\underline{x}$

$$\lim_{n\to\infty} \frac{1}{\text{Volume } (C^{(n)}(\underline{x}))} \int_{C^{(n)}(\underline{x})} \int_{-\infty}^{\infty} h(v) f(\underline{y}, v)\, dv\, d\underline{y} = \int_{-\infty}^{\infty} h(v) f(\underline{x}, v)\, dv, \tag{3.19}$$

by a well-known theorem on the differentiation of multiple Lebesgue integrals. A similar limit holds for the denominator of (3.18) so that

$$\begin{aligned} \lim_{n\to\infty} \varphi^{(n)}(\underline{x}) &= \frac{\int_{-\infty}^{\infty} h(v) f(\underline{x}, v)\, dv}{f(\underline{x})} \\ &= E(h(X_{1,r+1}) \mid X_1^{(r)} = \underline{x}) = \psi_r(\underline{x}), \end{aligned} \tag{3.20}$$

for almost all $\underline{x}$ such that $f(\underline{x}) > 0$. Similarly

$$\begin{aligned} \lim_{n\to\infty} \theta^{(n)}(\underline{x}) &= \lim_{n\to\infty} \frac{\int_{C^{(n)}(\underline{x})} \int_{-\infty}^{\infty} h^2(v) f(\underline{y}, v)\, dv\, d\underline{y}}{\int_{C^{(n)}(\underline{x})} f(\underline{y})\, d\underline{y}} \\ &= \frac{\int_{-\infty}^{\infty} h^2(v) f(\underline{x}, v)\, dv}{f(\underline{x})} \\ &= E(h^2(X_{1,r+1}) \mid X_1^{(r)} = \underline{x}), \end{aligned} \tag{3.21}$$

for almost all $\underline{x}$ such that $f(\underline{x}) > 0$, so that (3.15) and (3.16) are verified.

For $n = 1, 2, \cdots$, let

$$V^{(n)}(\underline{x}) = \begin{cases} \dfrac{1}{\bar{M}^{(n)}(\underline{x})}, & \bar{M}^{(n)}(\underline{x}) > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{3.22}$$

$$\xi^{(n)}(\underline{x}) = EV^{(n)}(\underline{x}), \tag{3.23}$$

and

$$P^{(n)}(\underline{x}) = \text{Prob } \{\bar{M}^{(n)}(\underline{x}) > 0\}. \tag{3.24}$$

For each n, $V^{(n)}(\underline{x})$, $\xi^{(n)}(\underline{x})$, and $P^{(n)}(\underline{x})$ are analogous respectively to $V_n(\underline{x})$, $\xi_n(\underline{x})$, and $P_n(\underline{x})$ of Sec. 2.

LEMMA 4. *For almost all x such that $f(x) > 0$*

$$\lim_{n\to\infty} \xi^{(n)}(x) = 0, \tag{3.25}$$

and

$$\lim_{n\to\infty} P^{(n)}(x) = 1. \tag{3.26}$$

Proof. Let

$$p^{(n)}(x) = \int_{C^{(n)}(x)} f(y)\, dy. \tag{3.27}$$

We have remarked that for almost all x

$$\lim_{n\to\infty} \frac{1}{\text{Volume } (C^{(n)}(x))} \int_{C^{(n)}(x)} f(y)\, dy = f(x). \tag{3.28}$$

By (3.8) and (3.9) we have

$$\text{Volume } (C^{(n)}(x)) = \frac{c^r}{n^{1-\delta}}, \tag{3.29}$$

so that whenever $f(x) > 0$ we may write

$$p^{(n)}(x) = \frac{c^r}{n^{1-\delta}} f(x)(1 + \epsilon_n(x)), \tag{3.30}$$

where $\lim_{n\to\infty} \epsilon_n(x) = 0$ for almost all x.

Referring to the upper bound (2.53) obtained for $\xi_n(x)$ in Sec. 2 and noting that $m(x)p^{(n)}(x)$ plays the same role as $p(x)$ and that $m(x) \geqq 1$, we see that

$$\xi^{(n)}(x) \leqq \frac{2}{np^{(n)}(x)} = \frac{2}{n^\delta c^r f(x)(1 + \epsilon_n(x))}, \tag{3.31}$$

for all x such that $f(x) > 0$. Hence

$$\lim_{n\to\infty} \xi^{(n)}(x) = 0, \tag{3.32}$$

for almost all x such that $f(x) > 0$ and (3.25) is verified. Now

$$P^{(n)}(x) = \text{Prob } \{\bar{M}^{(n)}(x) > 0\} \geqq 1 - [1 - p^{(n)}(x)]^n, \tag{3.33}$$

and since for all $u \leqq 1$, $e^{-u} \geqq 1 - u \geqq 0$, we may write

$$P^{(n)}(x) \geqq 1 - e^{-np^{(n)}(x)}. \tag{3.34}$$

Hence for almost all x such that $f(x) > 0$,

$$\liminf_{n\to\infty} P^{(n)}(x) \geqq 1 - \lim_{n\to\infty} e^{-n^\delta c^r f(x)(1+\epsilon_n(x))} = 1, \tag{3.35}$$

which implies (3.26).

We now prove a simple convergence lemma.

LEMMA 5. *If $(S, \mathcal{B}, \nu)$ is a measure space, $\{f_n\}$ and $\{g_n\}$ are sequences of non-negative integrable functions, f is an integrable function and g is a function such that*

$$\text{(i)}\quad \lim_{n\to\infty} f_n = f, \text{ a.e.}; \qquad \lim_{n\to\infty} g_n = g, \text{ a.e.};$$

$$\text{(ii)}\quad g_n \leqq f_n, \quad \text{all} \quad n; \tag{3.36}$$

$$\text{(iii)}\quad \limsup_{n\to\infty} \int f_n \, d\nu \leqq \int f \, d\nu;$$

then g is integrable and

$$\lim_{n\to\infty} \int g_n \, d\nu = \int g \, d\nu. \tag{3.37}$$

Proof. By (i), (ii) and Fatou's Lemma, g is integrable,

$$\liminf_{n\to\infty} \int g_n \, d\nu \geqq \int g \, d\nu, \tag{3.38}$$

and (noting (iii)),

$$\lim_{n\to\infty} \int f_n \, d\nu = \int f \, d\nu. \tag{3.39}$$

Furthermore,

$$\begin{aligned} \limsup_{n\to\infty} \int (f_n - g_n) \, d\nu &\leqq \limsup_{n\to\infty} \int f_n \, d\nu - \liminf_{n\to\infty} \int g_n \, d\nu \\ &\leqq \int (f - g) \, d\nu, \end{aligned} \tag{3.40}$$

and by (i) and (ii), $\lim_{n\to\infty} (f_n - g_n) = f - g$ and $(f_n - g_n) \geqq 0$, so that applying Fatou's Lemma again we obtain

$$\lim_{n\to\infty} \int (f_n - g_n) \, d\nu = \int f \, d\nu - \int g \, d\nu. \tag{3.41}$$

The desired result then follows from (3.39) and (3.41).

THEOREM 3. *If the a priori probability measure space $(\Omega, \mathcal{A}, \mu)$ is such that* (3.5) *is satisfied, then*

$$\lim_{n\to\infty} R(\psi_n) = R(\psi_\mu). \tag{3.42}$$

Proof. Since X and Λ are independent of $X_1, X_2, \cdots, X_n$, we have (as in Sec. 2)

$$E(\psi_n^2(X) \mid X = x) = E\psi_n^2(x), \tag{3.43}$$

and

$$E(\Lambda\psi_n^2(X) \mid X = x) = \psi_\mu(x) E\psi_n(x), \tag{3.44}$$

for almost all $\underline{x}$. Now applying Lemma 1 of Sec. 2 to $\psi_n(\underline{x})$ (mutatis mutandis) we obtain

$$E\psi_n(\underline{x}) = \varphi^{(n)}(\underline{x})P^{(n)}(\underline{x}), \tag{3.45}$$

and

$$E\psi_n^2(\underline{x}) = [\theta^{(n)}(\underline{x}) - \varphi^{(n)2}(\underline{x})]\xi^{(n)}(\underline{x}) + \varphi^{(n)2}(\underline{x})P^{(n)}(\underline{x}), \tag{3.46}$$

for almost all $\underline{x}$. Hence

$$E\psi_n^2(X) = E[\theta^{(n)}(X)\xi^{(n)}(X)] - E[\varphi^{(n)2}(X)\xi^{(n)}(X)] + E[\varphi^{(n)2}(X)P^{(n)}(X)], \tag{3.47}$$

and

$$E[\Delta\psi_n(X)] = E[\psi_n(X)\varphi^{(n)}(X)P^{(n)}(X)]. \tag{3.48}$$

Now in terms of the hypercubes $C_j^{(n)}$, $j = 1, 2, \cdots$, defined by (3.9) we may write

$$E\varphi^{(n)2}(X) = \sum_{j=1}^{\infty} \frac{\left[\int_{C_j^{(n)}} \int_{-\infty}^{\infty} h(v)f(\underline{u}, v)\, dv\, d\underline{u}\right]^2}{\int_{C_j^{(n)}} f(\underline{u})\, d\underline{u}}, \tag{3.49}$$

and

$$\begin{aligned} E\psi_n^2(X) &= \int \frac{\left[\int_{-\infty}^{\infty} h(v)f(\underline{u}, v)\, dv\right]^2}{f(\underline{u})}\, d\underline{u} \\ &= \sum_{j=1}^{\infty} \int_{C_j^{(n)}} \frac{\left[\int_{-\infty}^{\infty} h(v)f(\underline{u}, v)\, dv\right]^2}{f(\underline{u})}\, d\underline{u}, \end{aligned} \tag{3.50}$$

where it is understood that the ratios appearing in these expressions are to be replaced by zero whenever their denominators vanish. Furthermore, for each j we have

$$\left[\int_{C_j^{(n)}} \int_{-\infty}^{\infty} h(v)f(\underline{u}, v)dv\, d\underline{u}\right]^2 \leqq \int_{C_j^{(n)}} f(\underline{u})\, d\underline{u} \cdot \int_{C_j^{(n)}} \frac{\left[\int_{-\infty}^{\infty} h(v)f(\underline{u}, v)\, dv\right]^2}{f(\underline{u})}\, d\underline{u}, \tag{3.51}$$

by the Schwartz inequality. Hence

$$E\varphi^{(n)2}(X) \leqq E\psi_n^2(X), \qquad \text{all } n, \tag{3.52}$$

and since $0 \leqq P^{(n)}(\underline{x}) \leqq 1$,

$$E[\varphi^{(n)2}(X)P^{(n)}(X)] \leqq E\psi_n^2(X), \qquad \text{all } n. \tag{3.53}$$

Now by (3.15) of Lemma 3 and (3.26) of Lemma 4

$$\lim_{n\to\infty} \varphi^{(n)2}(\underline{x})P^{(n)}(\underline{x}) = \psi_\mu^2(\underline{x}), \tag{3.54}$$

for almost all $\underline{x}$ such that $f(\underline{x}) > 0$. Hence by Lemma 5 (with $f_n = g_n = \varphi^{(n)2}(\underline{x})P^{(n)}(\underline{x})$) we have

$$\lim_{n\to\infty} E[\varphi^{(n)2}(\underline{X})P^{(n)}(\underline{X})] = E\psi_\mu^2(\underline{X}). \tag{3.55}$$

Now

$$E\theta^{(n)}(\underline{X}) = Eh^2(X_{1,r+1}) = E\{E(h^2(X_{1,r+1}) \mid \underline{X}_1^{(r)})\}, \tag{3.56}$$

for all n, and by (3.16) of Lemma 3

$$\lim_{n\to\infty} \theta^{(n)}(\underline{x}) = E(h^2(X_{1,r+1}) \mid \underline{X}_1^{(r)} = \underline{x}), \tag{3.57}$$

for almost all $\underline{x}$ such that $f(\underline{x}) > 0$. Also, $0 \leqq \xi^{(n)}(\underline{x}) \leqq 1$ and by (3.25) of Lemma 4

$$\lim_{n\to\infty} \xi^{(n)}(\underline{x}) = 0, \tag{3.58}$$

for almost all $\underline{x}$ such that $f(\underline{x}) > 0$, so that by Lemma 5 (with $f_n = \theta^{(n)}(\underline{x})$ and $g_n = \theta^{(n)}(\underline{x})\xi^{(n)}(\underline{x})$) we have

$$\lim_{n\to\infty} E[\theta^{(n)}(\underline{X})\xi^{(n)}(\underline{X})] = 0. \tag{3.59}$$

Similarly, in view of (3.16) of Lemma 3 and (3.52),

$$\lim_{n\to\infty} E[\varphi^{(n)2}(\underline{X})\xi^{(n)}(\underline{X})] = 0. \tag{3.60}$$

Hence by (3.47), (3.55), (3.59), and (3.60)

$$\lim_{n\to\infty} E\psi_n^2(\underline{X}) = E\psi_\mu^2(\underline{X}). \tag{3.61}$$

Now for any fixed n and j the functions $\varphi^{(n)}(\underline{x})$ and $P^{(n)}(\underline{x})$ are constant for all $\underline{x} \in C_j^{(n)}$ and we may designate their values by $\varphi_j^{(n)}$ and $P_j^{(n)}$ respectively. Then by (3.48) we have for each n

$$\begin{aligned} E[\Delta\psi_n(\underline{X})] &= E[\psi_\mu(\underline{X})\varphi^{(n)}(\underline{X})P^{(n)}(\underline{X})] \\ &= \sum_{j=1}^{\infty} \varphi_j^{(n)}P_j^{(n)} \int_{C_j^{(n)}} \psi_\mu(\underline{x})f(\underline{x})\, d\underline{x} \\ &= \sum_{j=1}^{\infty} \varphi_j^{(n)}P_j^{(n)} \int_{C_j^{(n)}} \int_{-\infty}^{\infty} h(v)f(\underline{x}, v)\, dv\, d\underline{x} \\ &= \sum_{j=1}^{\infty} \varphi_j^{(n)2}P_j^{(n)} \int_{C_j^{(n)}} f(\underline{x})\, d\underline{x} \\ &= E[\varphi^{(n)2}(\underline{X})P^{(n)}(\underline{X})], \end{aligned} \tag{3.62}$$

so that by (3.55)

$$\lim_{n\to\infty} E[\Lambda\psi_n(X)] = E\psi_\mu^2(X). \tag{3.63}$$

Now

$$R(\psi_n) = E\Lambda^2 + E\psi_n^2(X) - 2E[\Lambda\psi_n(X)], \tag{3.64}$$

so that by (3.61) and (3.63) we have

$$\begin{aligned}\lim_{n\to\infty} R(\psi_n) &= E\Lambda^2 + \lim_{n\to\infty} E\psi_n^2(X) - 2\lim_{n\to\infty} E[\Lambda\psi_n(X)] \\ &= E\Lambda^2 - E\psi_\mu^2(X) = R(\psi_\mu),\end{aligned} \tag{3.65}$$

which was to be proved.

The estimation procedure introduced in this section contains an element of arbitrariness arising from the fact that the definition of the sequence of intervals $\{I_i^{(n)}\}$ involves the two constants c and δ whose values must be specified. The problem of the proper choice of c and δ will not, however, be considered further here.

The remarks made in Sec. 2 concerning various modifications of the estimator φ_n apply as well to the analogous modifications of the estimator ψ_n.

4. Hypothesis testing. The empirical Bayes estimation procedures introduced in the preceding sections may be applied to certain two-decision problems of the hypothesis-testing type. This is illustrated by the following two examples:

Example 1. (One-sided alternatives): Suppose that we wish to test a hypothesis about the value λ of the random variable Λ associated with the vector of observations X. In particular, suppose that we wish to test the hypothesis $H_0: \lambda < a$ versus the alternative hypothesis $H_1: \lambda \geqq a$. Let A_0 represent the action of accepting H_0 and let A_1 represent the action of accepting H_1. Then we may define a loss function L as follows:

$$L(A_i, \lambda) = \begin{cases} \max(0, \lambda - a), & i = 0, \\ -\min(0, \lambda - a), & i = 1. \end{cases} \tag{4.1}$$

In the decision theoretic framework this loss function is certainly no less reasonable than the classical zero-one loss function usually postulated for hypothesis-testing problems. Now for any decision function $\delta(x) = \text{Prob}\{A_1 \mid X = x\} =$ probability of rejecting H_0 when x is observed, the risk involved in using δ is given by

$$\begin{aligned} R(\delta) &= E\{\delta(X)L(A_1, \Lambda)\} + E\{[1 - \delta(X)]L(A_0, \Lambda)\} \\ &= EL(A_0, \Lambda) - E\{\delta(X)[L(A_0, \Lambda) - L(A_1, \Lambda)]\} \\ &= EL(A_0, \Lambda) - E\{\delta(X)[\Lambda - a]\} \\ &= EL(A_0, \Lambda) - E\{\delta(X)[E(\Lambda \mid X) - a]\}. \end{aligned} \tag{4.2}$$

Hence the Bayes decision function $\delta_\mu(x)$ minimizing $R(\delta)$ is

$$\delta_\mu(x) = \begin{cases} 1, & E(\Lambda \mid X = x) > a, \\ 0, & \text{otherwise.} \end{cases} \tag{4.3}$$

Now we have seen in the previous sections that when the a priori probability measure (and hence the joint distribution of Λ and X) is unknown we may still be able, under certain circumstances, to find an empirical Bayes estimator $\varphi_n(x)$ based on prior independent observations, such that

$$\lim_{n\to\infty} E[\varphi_n(x) - E(\Lambda \mid X = x)]^2 = 0, \tag{4.4}$$

for all x in some set S which is assigned probability 1 under the distribution of X. Now (4.4) implies that as $n \to \infty$,

$$\varphi_n(x) \to E(\Lambda \mid X = x), \qquad \text{in probability,} \tag{4.5}$$

for all $x \in S$, so that if we define the empirical Bayes decision function $\delta_n(x)$ by

$$\delta_n(x) = \begin{cases} 1, & \varphi_n(x) > a, \\ 0, & \text{otherwise,} \end{cases} \tag{4.6}$$

we will have

$$\lim_{n\to\infty} E\delta_n(x) = \lim_{n\to\infty} \text{Prob}\,\{\varphi_n(x) > a\} = \delta_\mu(x), \tag{4.7}$$

for all $x \in S$ with the possible exception of values of x for which $E(\Lambda \mid X = x) = a$. Hence

$$\begin{aligned} \lim_{n\to\infty} E\{\delta_n(X)[\Lambda - a] \mid X = x\} &= \lim_{n\to\infty} [E\delta_n(x)][E(\Lambda \mid X = x) - a] \\ &= \delta_\mu(x)[E(\Lambda \mid X = x) - a], \end{aligned} \tag{4.8}$$

for all $x \in S$. Also, since $0 \leqq \delta_n(x) \leqq 1$ for all values of x and n, we have

$$|[E\delta_n(x)][E(\Lambda \mid X = x) - a]| \leqq |E(\Lambda \mid X = x)| + |a|, \qquad \text{all } x, n. \tag{4.9}$$

Hence by (4.8), (4.9) and the Lebesgue Dominated Convergence Theorem we have

$$\lim_{n\to\infty} R(\delta_n) = R(\delta_\mu), \tag{4.10}$$

whenever the a priori probability measure is such that (4.5) holds and $E|\Lambda| < \infty$.

Example 2. (Two-sided alternatives): Suppose now that we wish to test the hypothesis $H_0^*: \lambda \in (a - b, a + b)$, $b > 0$, versus the alternative hypothesis $H_1^*: \lambda \notin (a - b, a + b)$, where, as before, λ is a value of the random variable

Λ which is associated with the vector of observations X. Let the loss function L^* be defined by

$$L^*(A_i, \lambda) = \begin{cases} \max(0, [(\lambda - a)^2 - b^2]), & i = 0, \\ -\min(0, [(\lambda - a)^2 - b^2]), & i = 1, \end{cases} \tag{4.11}$$

where, as before, A_i represents the action of accepting the hypothesis H_i^*, $i = 0, 1$. The graph of L^* is shown in Fig. 1. For any decision function $\delta(x) = \text{Prob}\{A_1 \mid X = x\}$, the risk is

$$\begin{aligned} R(\delta) &= EL^*(A_0, \Lambda) - E\{\delta(X)[L^*(A_0, \Lambda) - L^*(A_1, \Lambda)]\} \\ &= EL^*(A_0, \Lambda) - E\{\delta(X)[(\Lambda - a)^2 - b^2]\} \qquad (4.12) \\ &= EL^*(A_0, \Lambda) - E\{\delta(X)[E(\Lambda^2 \mid X) - 2aE(\Lambda \mid X) + a^2 - b^2]\}. \end{aligned}$$

Hence the Bayes decision function $\delta_\mu^*(x)$ minimizing $R(\delta)$ is given by

$$\delta_\mu^*(x) = \begin{cases} 1, & E(\Lambda^2 \mid X = x) - 2aE(\Lambda \mid X = x) > b^2 - a^2, \\ 0, & \text{otherwise.} \end{cases} \tag{4.13}$$

Now if the a priori probability measure is not known we may still be able, under certain circumstances, to find empirical Bayes estimators $\varphi_n^{(1)}(x)$ and $\varphi_n^{(2)}(x)$ based on prior independent observations, such that as $n \to \infty$,

$$\varphi_n^{(1)}(x) \to E(\Lambda \mid X = x), \qquad \text{in probability,} \tag{4.14}$$

and

$$\varphi_n^{(2)}(x) \to E(\Lambda^2 \mid X = x), \qquad \text{in probability,} \tag{4.15}$$

for all $x \in S$ where S, as before, is assigned probability 1 under the distribution of X. Then if we define the empirical Bayes decision function $\delta_n^*(x)$ by

$$\delta_n^*(x) = \begin{cases} 1, & \varphi_n^{(2)}(x) - 2a\varphi_n^{(1)}(x) > b^2 - a^2, \\ 0, & \text{otherwise,} \end{cases} \tag{4.16}$$

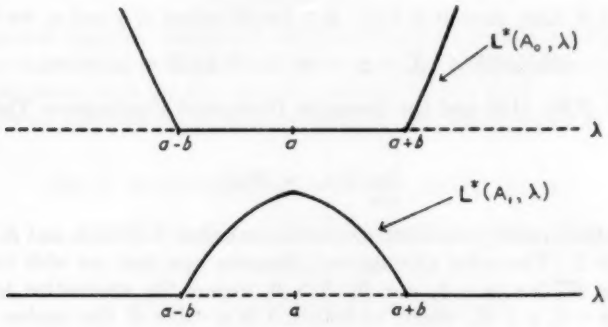

FIG. 1

we have (by the same argument as in Example 1),

$$\lim_{n\to\infty} R(\delta_n^*) = R(\delta_\mu^*), \tag{4.17}$$

for any a priori probability measure such that $E\Lambda^2 < \infty$ and (4.14) and (4.15) hold.

The existence of empirical Bayes estimators satisfying (4.14) follows directly (as in Example 1) from the results of the previous sections. We remark that if we assume that $Eh^4(X) < \infty$ in the cases treated in Secs. 2 and 3, then an empirical Bayes estimator satisfying (4.15) can be obtained whenever the number of components in the vector X_i exceeds the number in X by at least 2 for all i. To see this we observe that

$$E(h(X_{1,r+1})h(X_{1,r+2}) \mid X_1^{(r)} = x) = E(\Lambda^2 \mid X = x), \tag{4.18}$$

so that (in the notation of Sec. 2) if we let

$$\varphi_n^{(2)}(x) = \begin{cases} \dfrac{1}{\bar{M}_n(x)} \displaystyle\sum_{i=1}^{n} M_i(x)h(X_{i,r+1})h(X_{i,r+2}), & \bar{M}_n(x) > 0, \\ 0, \quad \text{otherwise}, \end{cases} \tag{4.19}$$

we can show (by arguments paralleling those of Sec. 2) that if $Eh^4(X) < \infty$ then

$$\lim_{n\to\infty} E[\varphi_n^{(2)}(x) - E(\Lambda^2 \mid X = x)]^2 = 0, \tag{4.20}$$

for all $x \in S$, which implies (4.15).

5. General remarks. The methods of this paper clearly may be modified to apply to compound Bayes decision problems where the component problems are of one of the types considered above and where the compound risk is the average of the component risks. Robbins has conjectured in [3] that empirical Bayes solutions of such compound problems will often lead to asymptotically subminimax solutions for the corresponding compound decision problems where no a priori probability measure is assumed to exist. We may surmise therefore that suitable modifications of the techniques given here are applicable to such problems.

6. Acknowledgements. I wish to express my thanks to Professor Herbert Robbins for proposing this investigation and for his helpful suggestions during its progress. I also wish to thank my wife, Maryann Haertter Johns, for performing the computations necessary for Tables I and II.

REFERENCES

[1] H. Robbins, "An empirical Bayes approach to statistics," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1956, pp. 157–163.

[2] M. Loève, *Probability Theory*, D. Van Nostrand, 1955.

[3] H. Robbins, "Asymptotically subminimax solutions of compound statistical decision problems," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1951, pp. 131–148.

ON QUEUES WITH POISSON ARRIVALS

By V. E. Beneš

Bell Telephone Laboratories, Incorporated

1. Summary. The system to be studied consists of a service unit and a queue of customers waiting to be served. Service-times of customers are independent, nonnegative variates with the common distribution $B(v)$ having a finite first moment b_1. Customers arrive in a Poisson process (see Feller [4], p. 364) of intensity λ; they form a queue and are served in order of arrival, with no defections from the queue. For previous work on this queueing system see for instance Pollaczek [11], Khintchine [9], Lindley [10], Kendall [7], [8], Smith [12], Bailey [1], and Takács [14].

Let $W(t)$ be the time a customer would have to wait if he arrived at t. The forward Kolmogorov equation for the distribution of $W(t)$ is solved in principle by the use of Laplace integrals, and $E\{\exp\{-sW(t)\}\}$ is determined in terms of $W(0)$ and the root of a possibly transcendental equation. It is shown that any analytic function of the root can be expanded in Lagrange's series, which provides a way of actually computing the transition probabilities of the process. Let z be the first zero of $W(t)$. A series for $E\{\exp\{-\tau z\}\}$ is obtained, and it is proved that $\text{pr}\{z < \infty\} = 1$ if and only if $\lambda b_1 \leqq 1$. From a functional relation between $E\{W(t)\}$ and $\text{pr}\{W(t) = 0\}$ the covariance function R of $W(t)$ is determined. If the service-time distribution $B(v)$ has a finite third moment, then R is absolutely integrable, and the spectral distribution of $W(t)$ is absolutely continuous.

2. The distributions of waiting-time and busy-time. Let $W(t)$ be the instantaneous waiting-time. That is, let $W(t)$ be the time that a customer arriving at t would have to wait before beginning his service. Evidently $W(t)$ jumps upward discontinuously every time someone arrives who has a nonzero service time. Otherwise $W(t)$ approaches 0 with slope -1 until it jumps again or reaches 0. At 0 it stays $=0$ until another jump occurs. The magnitudes of the jumps are the (independent) service-times of the customers arriving at the jumps. $W(t)$ is a continuous parameter Markov process of the mixed type considered in Feller [5]. Let $P(w, t) = \text{pr}\{W(t) \leqq w\}$. As shown in Takács [14], the forward Kolmogorov equation of the process is

$$\frac{\partial P(w, t)}{\partial t} = \frac{\partial P(w, t)}{\partial w} - \lambda P(w, t) + \lambda \int_{0-}^{w} P(w - v, t)\, dB(v), \tag{2.1}$$

and if $\varphi(s, t) = E\{\exp\{-sW(t)\}\}$ for $\text{Re}(s) \geqq 0$, we obtain

$$\frac{\partial \varphi(s, t)}{\partial t} = \varphi(s, t)[s - \lambda(1 - B^*)] - sP(0, t), \tag{2.2}$$

Received July 6, 1956; revised January 14, 1957.

where $B^*(s)$ is the Laplace-Stieltjes transform of $B(v)$. Let $\varphi^*(s, \tau)$ be the Laplace transform of φ with respect to t, and let $P^*(\tau)$ be that of $P(0, t)$.

Then the Kolmogorov equation implies

$$\varphi^* = \frac{\varphi(s, 0) - sP^*}{\tau - s + \lambda[1 - B^*]}. \tag{2.3}$$

By the busy-time, we mean the epoch of the first zero of $W(t)$ subsequent to 0, when $W(0)$ is any admissible starting point. The busy-time can be investigated in terms of a modified process which is like $W(t)$ except that it stays at 0 once it arrives there. Let z be the first zero in $t \geqq 0$, and define

$$Z(t) = W(t), \qquad \text{for } t \leqq z,$$

$$Z(t) = 0, \qquad \text{for } t > z.$$

Let $F(u, t) = \text{pr}\{Z(t) \leqq u\}$; the forward Kolmogorov equation for the process is

$$\frac{\partial F(u, t)}{\partial t} = \frac{\partial F(u, t)}{\partial u} - \lambda[F(u, t) - F(0, t)] + \lambda \int_{0-}^{u} F(u - v, t)\, dB(v) - \lambda F(0, t)B(u). \tag{2.4}$$

Let $\psi(s, t) = E\{\exp\{-sZ(t)\}\}$; let $\psi^*(s, \tau)$ be the Laplace transform (with respect to t) of ψ, and also let $F^*(\tau) = E\{e^{-\tau z}\}$, $f^* = \tau F^*$, for $\text{Re}(\tau) > 0$.

Then the Kolmogorov equation 2.4 yields

$$\psi^* = \frac{\psi(s, 0) - f^*[s - \lambda + \lambda B^*]}{\tau - s + \lambda[1 - B^*]}. \tag{2.5}$$

To solve for the unknown functions P^* and F^* we argue that the transforms φ^* and ψ^* must converge (cf. Bailey [1]) for $\text{Re}(s) > 0$ whenever $\text{Re}(\tau) > 0$, and that in this region zeros of $\tau - s + \lambda[1 - B^*]$ must coincide with zeros of the respective numerators. We show that there is an unique zero $\eta(\tau)$ of

$$\tau - s + \lambda[1 - B^*]$$

in $\text{Re}(s) > 0$ for $\text{Re}(\tau) > 0$. Choose a real δ such that $0 < \delta < \text{Re}(\tau)$, and a real $\epsilon > \text{Re}(\tau)$. Consider the line $\text{Re}(s) = \delta$, and the circle, with center at $\tau + \lambda$, defined by $|\tau - s + \lambda| = \lambda + \epsilon$. Define a contour C to be the circle when $\text{Re}(s) > \delta$, and to be the line when $\text{Re}(s) = \delta$. On the circle, we have the inequality

$$|\tau - s + \lambda| = \lambda + \epsilon > \lambda \geqq |\lambda B^*(s)|.$$

And on the line $\text{Re}(s) = \delta$:

$$|\tau - s + \lambda| \geqq \text{Re}(\tau) - \delta + \lambda > \lambda \geqq |\lambda B^*(s)|,$$

so that the inequality

$$|\tau - s + \lambda| > |\lambda B^*(s)|$$

holds over the whole contour C. Now $\tau - s + \lambda$ has no zeros on $\mathrm{Re}(s) = \delta$, nor any on the circle $|\tau - s + \lambda| = \lambda + \epsilon$; and $B^*(s)$ is single-valued and analytic in $\mathrm{Re}(s) > 0$. So by Rouché's theorem we conclude that $\tau - s + \lambda$ and $\tau - s + \lambda - \lambda B^*(s)$ have the same number of zeros in $\mathrm{Re}(s) > 0$, namely one, because δ can be made arbitrarily small, and ϵ arbitrarily large.

It follows that, with $\eta = \eta(\tau)$,

$$P^*(\tau) = \frac{\varphi(\eta, 0)}{\eta} = \frac{E\{e^{-\eta W(0)}\}}{\eta}, \tag{2.6}$$

$$F^*(\tau) = \psi(\eta, 0) = E\{e^{-\eta z(0)}\}. \tag{2.7}$$

In the proof above we saw that $|\tau - s + \lambda| > |\lambda B^*(s)|$, if s is on the contour C. So by Lagrange's expansion (p. 133 of [16]), for any function Γ analytic on and inside C, we have

$$\Gamma(\eta) = \Gamma(\tau + \lambda) + \sum_{n=1}^{\infty} \frac{(-\lambda)^n}{n!} \frac{d^{n-1}}{ds^{n-1}} \left[\frac{d\Gamma}{ds} (B^*(s))^n \right]_{s=\tau+\lambda}; \tag{2.8}$$

this expansion is valid for $\mathrm{Re}(\tau) > 0$, and provides a way of actually evaluating P^* and F^*. Except for the matter of inverting transforms, the solutions for the distributions of $W(t)$, $Z(t)$, and z are complete. It is easy to see that both φ^* and ψ^* can be inverted explicitly as Laplace transforms with respect to τ and give rise to exponential functions which, together with P^* and F^*, determine φ and ψ in terms of initial conditions.

The results of this section may be collected in the following statements.

THEOREM 1. *The function $\varphi(s, t) = E\{\exp\{-sW(t)\}\}$ is determined as the solution to Eq.* (2.2) *by the conditions*

$$\varphi(s, t) = \varphi(s, 0) \exp\{st - \lambda t[1 - B^*]\}$$

$$- s \int_0^t P(0, t - y) \exp\{sy - \lambda y[1 - B^*]\}\, dy, \tag{i}$$

$$P^*(\tau) = \int_0^\infty e^{-\tau t} P(0, t)\, dt = \eta^{-1}\varphi(\eta, 0) = \eta^{-1} E\{e^{-\eta W(0)}\}, \tag{ii}$$

where $\eta = \eta(\tau)$ is the unique root of $\tau - \eta + \lambda = \lambda B^*(\eta)$ in the right half-plane.

THEOREM 2. *The function $\psi(s, t) = E\{\exp\{-sZ(t)\}\}$ is determined by the conditions*

$$\psi(s, t) = \psi(s, 0) \exp\{st - \lambda t[1 - B^*]\}$$

$$- [s - \lambda + \lambda B^*] \int_0^t F(0, t - y) \exp\{sy - \lambda y[1 - B^*]\}\, dy, \tag{i}$$

(ii) $$\int_0^\infty e^{-\tau t}F(0, t)\,dt = \tau\psi(\eta, 0) = \tau E\{e^{-\eta z(0)}\},$$

where η is as in Theorem 1.

THEOREM 3. *If Γ is analytic in the open right half-plane and η is as in Theorem* 1, *then $\Gamma(\eta)$ may be expanded by Lagrange's series.*

3. The probability that z is finite. Since $F^* = E\{e^{-\tau z}\} = \psi(\eta, 0)$ where η is the root of $\tau - \eta + \lambda[1 - B^*(\eta)] = 0$, it seems natural to consider Tauberian and Abelian theorems in an effort to find the probability that z is finite, and to ascertain the existence of moments. We therefore turn attention to the behavior of η as $\tau \to 0$ along the real axis. There is an advantage to considering, instead of η, the linear function ξ of it defined by $\eta = \lambda(1 - \xi)$. Set $K(s) = B^*(\lambda - \lambda s)$, so that K generates a discrete probability distribution with mean λb_1 ; the equation for η may now be rewritten

$$\frac{\tau}{\lambda} + \xi = K(\xi),$$

and this fact suggests that as $\tau \to 0$ along the real axis, ξ approaches a root of the familiar equation from branching-process theory, $\xi = K(\xi)$. Let ζ be the least nonnegative real root of $\xi = K(\xi)$; for properties of ζ see Feller [4] or Harris [6]. We now show that $\xi \to \zeta$ as $\tau \to 0$ along the real axis.

If τ is real then so is ξ; for if not, then η is conjugate and not unique. Also, if $\tau > 0$, then $\xi < \zeta$, because $\tau > 0$ implies $K(\xi) > \xi$, and in $\xi < 1$ this is possible only if $\xi < \zeta$, since $K(0) > 0$ and $\xi = K(\xi)$ has at most two roots in (0, 1), one of them being at 1. To show that $0 < \tau < \tau'$ implies $\xi(\tau) > \xi(\tau')$, write $\xi = \xi(\tau)$, $\xi' = \xi(\tau')$. Then the hypothesis and $\xi < \zeta$, $\xi' < \zeta$ imply

$$K(\xi) - K(\xi') < \xi - \xi'.$$

Now $K'(y)$ is steadily increasing in $0 < y < 1$, so if for some u we have both $u < \zeta$ and $K'(u) > 1$, then $K(u) > u$ and $K(1) > 1$; so $K'(y) \leqq 1$ for $y < \zeta$. Now if $\xi \leqq \xi'$, this would imply

$$K(\xi') - K(\xi) \leqq \xi' - \xi,$$

which is impossible. It remains to show that given $u < \zeta$, there exists a $\tau > 0$ such that $\xi(\tau) > u$. The equation $x = \lambda[K(u) - u]$ uniquely determines an $x > 0$, and for this x we must have $\xi(x) = u$, or else $\tau - \eta + \lambda[1 - B^*(\eta)] = 0$ does not have a unique root η. If now $0 < \tau < x$, then $\xi(\tau) > \xi(x) = u$, as was to be proved.

It follows that as $\tau \to 0$ along the real axis, $\eta \to 0$ or $\lambda(1 - \zeta)$ according as $\lambda b_1 \leqq 1$ or $\lambda b_1 > 1$. We are now in a position to prove that

$$\text{pr}\{z < \infty\} = \lim_{t\to\infty} F(0, t) = E\{\exp\{\lambda(\zeta - 1)Z(0)\}\}.$$

As $\tau \to 0$ along the real axis, the continuity of ψ yields

$$\begin{aligned} F^*(\tau) &\to \psi(\lambda(1 - \zeta), 0) \\ &\to E\{\exp\{\lambda(\zeta - 1)Z(0)\}\}. \end{aligned}$$

Since $F(0, t)$ is nondecreasing, so is

$$\int_0^t xF(0, dx);$$

thus $F(0, t)$ and F^* satisfy the hypothesis of Theorem 4.5 of Widder [15]. This proves

THEOREM 4. *The probability that the first zero z of $W(t)$ is finite is*

$$\text{pr}\{z < \infty\} = \lim_{\tau \to 0} F^*(\tau) = E\{\exp\{\lambda(\zeta - 1)Z(0)\}\}.$$

This limit is 1 if $\lambda b_1 \leqq 1$, and is <1 if $\lambda b_1 > 1$.

A discussion similar to the above has been given by Takács [14] for the case $\varphi(s, 0) = B^*(s)$, and this case is also treated by Kendall [7]. We mention in addition the following results, provable by simple Abelian arguments: If $\lambda b_1 < 1$ and $E\{Z(0)\} < \infty$, then

$$E\{z\} = \frac{E\{Z(0)\}}{1 - \lambda b_1};$$

if $\lambda b_1 \geqq 1$, or $E\{Z(0)\} = \infty$, then $E\{z\} = \infty$.

4. The expectation of $W(t)$.

THEOREM 5. *If $E\{W(0)\} < \infty$, then $E\{W(t)\}$ exists for $t > 0$ and is given by*

$$E\{W(t)\} = E\{W(0)\} + \int_0^t [P(0, u) - 1 + \lambda b_1]\, du. \tag{4.1}$$

A result similar to this appears in Clarke [2]. To prove (4.1), we differentiate (2.2) with respect to s, and let $s \to 0$. From (4.1) we see that if $\lambda b_1 > 1$, then $d/dt\, E\{W(t)\}$ is positive and bounded away from 0, so that $E\{W(t)\}$ increases indefinitely. Let $M^*(\tau)$ be the Laplace transform with respect to t of $E\{W(t)\}$; then (4.1) implies

$$\begin{aligned} M^*(\tau) &= \frac{E\{W(0)\} + P^*}{\tau} - \frac{1 - \lambda b_1}{\tau^2} \\ &= \frac{E\{W(0)\} + \eta^{-1}E\{\exp\{-\eta W(0)\}\}}{\tau} - \frac{1 - \lambda b_1}{\tau^2}. \end{aligned}$$

From this it can be shown that if $B(v)$ has a finite second moment b_2, and $\lambda b_1 < 1$, then

$$\lim_{t \to \infty} E\{W(t)\} = \lim_{\tau \to 0} \tau M^* = \frac{\lambda b_2}{2(1 - \lambda b_1)}.$$

5. The stationary distribution. We call an initial distribution $P(w, 0)$ of $W(0)$ stationary if it is invariant under the transition probabilities for $W(t)$, that is, when $P(w, t) = P(w, 0)$ for all t and w. Let $A(w)$ be the distribution whose Laplace-Stieltjes transform is given by the Pollaczek-Khintchine formula

$$A^*(s) = \frac{s(1 - \lambda b_1)}{s - \lambda[1 - B^*]}, \qquad \lambda b_1 < 1.$$

We show that $A(w)$ is the unique stationary distribution. From (2.3) we see that a given $P(w, 0)$ is stationary if and only if the corresponding $\varphi(s, 0)$ satisfies

$$\frac{\varphi(s, 0)}{\tau} = \frac{\varphi(s, 0) - [s\varphi(\eta, 0)]/\eta}{\tau - s + \lambda[1 - B^*(s)]},$$

$$P^* = \frac{\varphi(\eta, 0)}{\eta} = \frac{P(0, 0)}{\tau}.$$

These imply

$$\varphi(s, 0) = \frac{sP(0, 0)}{s - \lambda[1 - B^*]},$$

and a simple Abelian argument proves $P(0, 0) = 1 - \lambda b_1$. This shows that $A(w)$ is unique.

To invert the transform $A^*(s)$ explicitly, we write it as

$$\frac{1 - \lambda b_1}{1 - \lambda b_1(1 - B^*)/b_1 s}$$

and notice that since $B(v)$ is the distribution of a nonnegative random variable with mean $0 < b_1 < \infty$, therefore

$$\frac{1 - B^*}{b_1 s}$$

is the Laplace transform of the density function

$$h(v) = \frac{U(v) - B(v)}{b_1},$$

where $U(v)$ is the unit step at 0. Define

$$H_0(w) = U(w),$$

$$H_{n+1}(w) = \int_0^w H_n(w - v)h(v)\, dv.$$

Then $A^*(s)$ may be inverted, and we have proved

THEOREM 6. *$A(w)$ is the unique stationary distribution of $W(0)$.* It may be written as

$$A(w) = (1 - \lambda b_1) \sum_{n=0}^{\infty} (\lambda b_1)^n H_n(w),$$

which shows that $A(w)$ is decomposable into a single step of magnitude $1 - \lambda b_1$ at 0 and an absolutely continuous portion, and that the equilibrium solution of Pollaczek and Khintchine has the form of a compound geometric distribution, i.e.,

$$W(\infty) = \sum_{i=0}^{k} x_i,$$

where the x's are mutually independent with the common density $h(v)$, and $\mathrm{pr}\{k = n\} = (1 - \lambda b_1)(\lambda b_1)^n$.

6. The covariance function and the spectral distribution. In this last section we assume that $\lambda b_1 < 1$, and that $W(0)$ has the stationary distribution $A(w)$. Let $E\{W^n(0)\} = a_n$, when this exists. The covariance function $R(t)$ of the process is

$$R(t) = \int_{0-}^{\infty} wE\{W(t) \mid W(0) = w\}\, dA(w) - a_1^2, \tag{6.1}$$

and the Laplace transform of $R(t)$ is

$$R^*(\tau) = \int_{0-}^{\infty} \left[\frac{w^2 - wa_1}{\tau} + \frac{we^{-\eta w}}{\eta\tau} - \frac{w(1 - \lambda b_1)}{\tau^2}\right] dA(w). \tag{6.2}$$

Intuitively one might expect that in view of the Poisson arrival process and the independent service-times, the $W(t)$ process would have no periodic components, and thus a smooth spectral distribution. That this is so under weak conditions is a consequence of the following:

THEOREM 7. *If $\lambda b_1 < 1$, and $B(v)$ has a finite moment b_3 of the third order, then*

$$\int_0^{\infty} |R(t)|\, dt < \infty. \tag{6.3}$$

To prove that $R(t)$ is $L_1(0, \infty)$, we show that $R^*(\tau)$ is the Laplace-Stieltjes transform of an absolutely continuous (AC) function of bounded total variation (BTV). We make use of the following result: Let u be a nonnegative random variable such that $E\{u\}$ and $E\{u^2\}$ both exist; then

$$E\{e^{-\tau y}\} = \frac{1 - E\{e^{-\tau u}\}}{\tau E\{u\}}$$

defines an unique variate $y \geqq 0$ such that distr $\{y\}$ is AC and $E\{y\} = E\{u^2\} / 2E\{u\}$.

By differentiating $A^*(s)$ successively, it can be verified that if b_3 exists, so do a_1 and a_2, and $a_1 = \lambda b_2 / 2(1 - \lambda b_1)$. Let $N^* = \lambda B^*(\eta) / (\tau + \lambda)$, and let

$$T_1^* = \frac{1 - [(\lambda e^{-\eta w})/(\tau + \lambda)]}{[\tau w/(1 - \lambda b_1)] + \tau\lambda^{-1}},$$

$$T_2^* = \frac{1 - [\lambda(1 - \lambda b_1)(1 - N^*)]/\tau}{[\tau a_1/(1 - \lambda b_1)] + \tau\lambda^{-1}},$$

By use of $\eta = \tau + \lambda - \lambda B^*(\eta)$ and algebra we can write the integrand of (6.2) as

$$w\lambda^{-1}[w(1 - \lambda b_1)^{-1} + \lambda^{-1}][\lambda(1 - \lambda b_1)(1 - N^*)\tau^{-1} - T_1^*](1 - N^*)^{-1}$$
$$-w\lambda^{-1}[a_1(1 - \lambda b_1)^{-1} + \lambda^{-1}][\lambda(1 - \lambda b_1)(1 - N^*)\tau^{-1} - T_2^*](1 - N^*)^{-1}.$$

By Taylor series arguments and repeated use of the result stated earlier, it can be shown that if b_3 exists, then each of T_1^*, T_2^*, and N^* is $E\{\exp\{-\tau y\}\}$ for some

suitable $y \geqq 0$, such that $E\{y\} < \infty$ and distr $\{y\}$ is AC. It follows from Lemma 5 of Smith [13] that for each w, the integrand of (6.2) is the Laplace-Stieltjes transform of an AC function of BTV. Therefore $R^*(\tau)$ is also.

From (6.3), and from the remarks on p. 522 of Doob [3], it follows that if $\lambda b_1 < 1$, and $B(v)$ has a finite moment b_3 of third order, then $W(t)$ has an absolutely continuous spectral distribution. The associated spectral density $g(x)$ is continuous and is given by

$$g(x) = 4 \int_0^\infty R(t) \cos 2\pi x t \, dt = 4 \, Re\{R^*(2\pi i x)\},$$

since R^* is well defined along the imaginary axis.

REFERENCES

[1] N. T. J. Bailey, "A continuous time treatment of a simple queue using generating functions," *J. Roy. Stat. Soc.*, Series B, Vol. 15 (1954), pp. 288–291.

[2] A. B. Clarke, "A waiting line process of Markov type," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 452–459.

[3] J. L. Doob, *Stochastic Processes*, John Wiley & Sons, Inc., New York, 1953.

[4] W. Feller, *An Introduction to Probability Theory and Its Applications*, John Wiley & Sons, Inc., New York, 1950.

[5] W. Feller, "Zur Theorie der Stochastischen Prozesse," *Math. Ann.*, Vol. 113 (1936), pp. 113–160.

[6] T. E. Harris, "Branching processes," *Ann. Math. Stat.*, Vol. 19 (1948), pp. 474–494.

[7] D. G. Kendall, "Some problems in the theory of queues," *J. Roy. Stat. Soc.*, Series B, Vol. 13 (1951), pp. 151–173 and 184–185.

[8] D. G. Kendall, "Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain," *Ann. Math. Stat.*, Vol. 24 (1953), pp. 338–354.

[9] A. Khintchine, "Matematischeskaya teoriya statsionarnoi ocheredi," *Matematicheskii sbornik*, Vol. 39 (1932), pp. 73–84.

[10] D. V. Lindley, "The theory of queues with a single server," *Proc. Cambridge Philos. Soc.*, Vol. 48 (1952), pp. 277–289.

[11] F. Pollaczek, "Uber eine Aufgabe der Wahrscheinlichkeitsrechnung," *Math. Zeit.*, Vol. 32 (1930), pp. 64–100 and 729–850.

[12] W. L. Smith, "On the distribution of queueing times," *Proc. Cambridge Philos. Soc.*, Vol. 49 (1953), pp. 449–461.

[13] W. L. Smith, "Asymptotic renewal theorems," *Proc. Roy. Soc. Edinburgh, Sec. A*, Vol. 64 (1954), pp. 9–48.

[14] L. Takács, "Investigation of waiting time problems by reduction to Markov processes," *Acta Math.* (Budapest) Vol. 6 (1955), pp. 101–129.

[15] D. Widder, *The Laplace Transform*, Princeton, 1941.

[16] E. Whittaker and G. N. Watson, *Modern Analysis*, Cambridge, 1946.

IDEMPOTENT MATRICES AND QUADRATIC FORMS IN THE GENERAL LINEAR HYPOTHESIS

By Franklin A. Graybill and George Marsaglia

Oklahoma A. and M. College

1. Introduction. The important role that idempotent matrices play in the general linear hypothesis theory has long been recognized ([1], [2]), but their usefulness seems not to have been fully exploited. The purpose of this paper is to state and prove some theorems about idempotent matrices and to point out how they might be used to advantage in linear hypothesis theory.

2. Notation and Definitions. Throughout this paper an idempotent matrix will mean a symmetric matrix A such that $AA = A$ (for the sake of brevity we will use the word idempotent matrix to indicate a symmetric idempotent matrix unless specifically stated otherwise). The theorems will not necessarily hold for nonsymmetric idempotent matrices. The statement: Y is distributed as $N_p(\mu,V)$, will mean that a $(p \times 1)$ random vector Y has the p-variate normal distribution whose mean is the $(p \times 1)$ vector, μ, and whose covariance matrix is the positive definite symmetric matrix, V. The statement: u is distributed as $\chi^2(n)$ will mean that a scalar random variable u has the Chi-square distribution with n degrees of freedom, and the statement: v is distributed as $\chi'^2(n, \lambda)$ will mean that the scalar random variable v is distributed as the noncentral Chi-square distribution with n degrees of freedom and with noncentrality, λ. The frequency function of v is ([3])

$$f(v) = \sum_{i=0}^{\infty} \frac{e^{-\lambda}\lambda^i}{i!} \cdot \frac{v^{n+2i-2/2}e^{-(v/2)}}{2^{n+2i/2}\Gamma\left(\frac{n+2i}{2}\right)}, \qquad 0 \leqq v < \infty.$$

If $\lambda = 0$, then the noncentral Chi-square distribution degenerates into the central Chi-square distribution.

A' will indicate the transpose of the matrix A, and A^{-1} will indicate the inverse. I_p will indicate the $(p \times p)$ identity matrix and φ will indicate a null matrix. Below is a list of well-known theorems which will be needed in the succeeding sections.

Theorem A. *If A is an $(n \times n)$ symmetric matrix of rank p, then a necessary and sufficient condition that A is idempotent is that each of p of the characteristic roots of A is equal to unity and the remaining $(n - p)$ characteristic roots are equal to zero.*

Theorem B. *If A is an idempotent matrix, then the rank of A equals the trace of A.*

Theorem C. *The only nonsingular idempotent matrix is the identity matrix.*

Theorem D. *If A is an $(n \times n)$ idempotent matrix of rank p such that $p < n$ $(p = n)$, then A is a positive semidefinite matrix (positive definite matrix).*

Received March 2, 1956; revised January 17, 1957.

THEOREM E. *If A is an idempotent matrix whose ith diagonal element is equal to zero, then every element in the ith row and ith column of A is equal to zero.*

THEOREM F. *If Y is distributed as $N_n(\mu, I_n)$, then $v = Y'Y$ is distributed as $\chi'^2(n, \lambda)$, where $\lambda = \frac{1}{2}\mu'\mu$.*

THEOREM G. *In Theorem F, the moment generating function of v is*

$$m_v(\theta) = (1 - 2\theta)^{-n/2} e^{-\lambda + \lambda(1-2\theta)^{-1}}.$$

THEOREM H. *If Y is distributed as $N_n(\mu, I_n)$, then a necessary and sufficient condition that $Y'B_1Y$, $Y'B_2Y$, $\cdots$, $Y'B_kY$ be jointly independent is that $B_iB_j = \varphi$ for all $i \neq j$.*

THEOREM J. *If $B_1, B_2, \cdots, B_k$ are a set of $(n \times n)$ symmetric matrices, then a necessary and sufficient condition that there exists an orthogonal matrix, P, such that $P'B_1P$, $P'B_2P$, $\cdots$, $P'B_kP$ are each diagonal is that $B_iB_j = B_jB_i$ for all i and j.*

THEOREM K. *Let $B_1, B_2, \cdots, B_m$ be a collection of $(n \times n)$ symmetric matrices such that $\sum_{i=1}^{m} B_i = I_n$. Then any one of the conditions K_1, K_2, K_3 is necessary and sufficient for the remaining two.*

K_1 : *Each B_i is an idempotent matrix.*

K_2 : *$B_iB_j = \varphi$ for all $i \neq j$.*

K_3 : *$\sum_{i=1}^{m} n_i = n$ where n_i is the rank of B_i.*

THEOREM L. *If v is distributed as $\chi'^2(n, \lambda)$ and w is distributed as $\chi^2(m)$, and if v and w are independent, then $u = (v/w)\cdot(m/n)$ is distributed as $F'(n, m, \lambda)$ where $F'(n, m, \lambda)$ refers to the noncentral F distribution with n degrees of freedom for numerator and m degrees of freedom for the denominator and noncentrality, λ. The functional form is*

$$f(u) = \sum_{i=0}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!} \frac{\Gamma\left(\frac{m+n+2i}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n+2i}{2}\right)} \left(\frac{n}{m}\right)^{n/2} \frac{u^{i+(n/2)-1}}{\left(1+\frac{n}{m}u\right)^{i+(n+m)/2}}.$$

This reduces to Snedecor's F if and only if $\lambda = 0$.

3. Theory. Let an observation vector, Y, be distributed as $N_n(X\beta, \sigma^2 I_n)$, where X is an $(n \times p)$ $(p < n)$ matrix with known elements and rank p, β is a $(p \times 1)$ vector of unknown parameters, and σ^2 is an unknown scalar. Y is often assumed to have this structure in models which are referred to as multiple regression models and in linear models used in the theory of experimental designs. In these models it is often desired to test hypotheses about elements of the vector β. The technique often employed to devise test functions is the technique of analysis of variance. The procedure is to partition the total sum of squares $Y'Y$ of the observation vector, Y, into quadratic forms such that

$$Y'Y = \sum_{i=1}^{k} Y'A_iY \tag{1}$$

and use Cochran's theorem ([5]) to ascertain the independence and distribution of the quantities $Y'A_iY$. This process is quite well known and will not be explained here except to say that to use Cochran's theorem it is necessary to be able to judge the rank of the matrices A_i. It has been pointed out ([2]) that in certain cases, finding the rank of the matrices A_i and using Cochran's theorem is equivalent to showing that $A_iA_j = \varphi$ for all $i \neq j$, or to showing that each A_i is an idempotent matrix. In many cases it is easier to show that a matrix is idempotent than it is to find the rank of the matrix. Therefore, we will prove some theorems which are new, and which enable us to determine the distribution of the quadratic forms in equations similar to (1) without having to find the rank of the A_i.

The first theorem which we shall prove is an algebraic theorem about symmetric matrices which is useful in developing theorems concerning the distribution of quadratic forms.

THEOREM 1. *Let $A_1, A_2, \cdots, A_m$ be a collection of $n \times n$ symmetric matrices where the rank of A_i is p_i, and let $A = \sum_{i=1}^{m} A_i$ where the rank of A is p. Consider the four conditions*:

C_1. *Each A_i is an idempotent matrix.*

C_2. *$A_iA_j = \varphi$ for all $i \neq j$.*

C_3. *A is an idempotent matrix.*

C_4. *$p = \sum_{i=1}^{m} p_i$; i.e., the rank of the sum of the A_i equals the sum of the ranks of the A_i.*

The following are true:

(a) *Any two of the three conditions C_1, C_2, C_3 imply all four of the conditions C_1, C_2, C_3, C_4.*

(b) *Conditions C_2 and C_4 imply C_1 and C_2.*

Proof. We will first prove (a). To do this we will show that any two of the conditions C_1, C_2, C_3 imply the remaining one in the set C_1, C_2, C_3, and then show that the three conditions C_1, C_2, and C_3 imply C_4. We might point out that if $A = I$ then this is essentially the theorem which Craig and Hotelling proved ([2], [4]).

Suppose C_1 and C_3 are given. Since A is given to be idempotent of rank p, there exists an orthogonal transformation P such that

$$P'AP = \begin{pmatrix} I_p & \varphi \\ \varphi & \varphi \end{pmatrix}.$$

Thus we have

$$P'AP = \begin{pmatrix} I_p & \varphi \\ \varphi & \varphi \end{pmatrix} = \sum_{i=1}^{m} P'A_iP.$$

Since A_i is idempotent, $P'A_iP$ is also idempotent, and by Theorems D and E, the last $(n - p)$ diagonal elements of each $P'A_iP$ must be zero. This is true since by Theorem D the diagonal elements of an idempotent matrix are non-

negative and since any one of the last $(n - p)$ diagonal elements when summed over the m matrices is zero, each of the last $(n - p)$ diagonal elements must be zero. Then by Theorem E the last $(n - p)$ rows and $(n - p)$ columns of each $P'A_iP$ must be zero. Thus we can write

$$P'A_iP = \begin{pmatrix} B_i & \varphi \\ \varphi & \varphi \end{pmatrix}.$$

Extracting the $(p \times p)$ matrix in the upper left-hand corner of

$$P'AP = \sum_{i=1}^{m} P'A_iP$$

we have $I_p = \sum_{i=1}^{m} B_i$, where the B_i are idempotent of rank p_i. Theorem K implies that $B_iB_j = \varphi$ for $i \neq j$; therefore $A_iA_j = \varphi$ if $i \neq j$, and the proof is complete that C_1 and C_3 imply C_2.

Now suppose C_1 and C_2 are given. We have

$$AA = \left(\sum_{i=1}^{m} A_i\right)^2 = \sum_{i=1}^{m} A_i^2 + \sum_{i \neq j} A_iA_j = \sum_{i=1}^{m} A_i = A.$$

Thus we have shown that the sum is idempotent and C_3 is satisfied.

Now suppose C_2 and C_3 are given. By Theorem J there exists an orthogonal matrix P such that $P'A_1P, P'A_2P, \cdots, P'A_mP$ are each diagonal (since $A_iA_j = A_jA_i = \varphi$), and since the sum of diagonal matrices is a diagonal matrix it also follows that $P'AP$ is diagonal. By C_2 it follows that $P'A_iPP'A_jP = \varphi$ for all $i \neq j$.

It follows, therefore, that $P'A_iP$ is idempotent, and hence A_i is idempotent for all i, and the proof is complete.

We will now show that C_1, C_2, and C_3 imply C_4. If the three conditions C_1, C_2, and C_3 are true, then this implies that there exists an orthogonal matrix P such that the following are true:

$P'AP = \begin{pmatrix} I_p & \varphi \\ \varphi & \varphi \end{pmatrix}$; $P'A_iP$ are each diagonal matrices with p_i (the rank of A_i) ones on the diagonal and $(n - p_i)$ zeros on the diagonal. Thus since

$$\sum_{i=1}^{m} P'A_iP = \begin{pmatrix} I_p & \varphi \\ \varphi & \varphi \end{pmatrix},$$

it is quite clear that the total number of ones on the diagonal of $P'A_iP$ $(i = 1, 2, \cdots, m)$ is equal to p and the result follows.

We will now prove (b). Since A is given as idempotent, there exists an orthogonal matrix P such that $P'AP = \begin{pmatrix} I_p & \varphi \\ \varphi & \varphi \end{pmatrix}$. Applying this transformation to the A_i gives $P'A_iP = M_i$ and M_i has rank p_i. Partition M_i such that

$$M_i = \begin{pmatrix} B_i & C_i' \\ C_i & D_i \end{pmatrix}$$

where B_i is a $p \times p$ symmetric matrix. Since $\sum_{i=1}^{m} M_i = \begin{pmatrix} I_p & \varphi \\ \varphi & \varphi \end{pmatrix}$ we have $\sum_{i=1}^{m} B_i = I_p$. Clearly the rank of B_i must be less than or equal to the rank of M_i. Therefore, let the rank of B_i equal $p_i - k_i$ where $k_i \geqq 0$. But the rank of the sum of matrices is less than or equal to the sum of the ranks, hence

$$\sum_{i=1}^{m} (p_i - k_i) \geqq p.$$

This gives $-\sum_{i=1}^{m} k_i \geqq 0$, so $k_i = 0$ for $i = 1, 2, \cdots, m$, and the rank of B_i is equal to p_i. Applying Theorem K to the equation $\sum B_i = I_p$ it follows that B_i is idempotent ($i = 1, 2, \cdots, m$) and $B_iB_j = \varphi$ for all $i \neq j$. By Theorem J we know that there exists an orthogonal matrix Q such that $Q'B_iQ$ is diagonal for $i = 1, 2, \cdots, m$. Let $Q'B_iQ = E_i$ where E_i is a $p \times p$ diagonal matrix with p_i diagonal elements equal to unity and the remaining diagonal elements equal to zero. Also, it follows that $\sum_{i=1}^{m} E_i = I_p$, so there is exactly one matrix in the set $E_1, E_2, \cdots, E_m$ whose tth diagonal element (for any $t = 1, 2, \cdots, p$) is equal to unity. All the remaining E_i have the tth diagonal element equal to zero. Since Q is orthogonal we know that

$$\underset{n\times n}{R} = \begin{pmatrix} Q & \varphi \\ \varphi & I_{n-p} \end{pmatrix}$$

is also orthogonal. Using this transformation on the equation

$$\sum_{i=1}^{m} M_i = \begin{pmatrix} I_p & \varphi \\ \varphi & \varphi \end{pmatrix}$$

gives

$$R'M_iR = \begin{pmatrix} E_i & F_i' \\ F_i & G_i \end{pmatrix}$$

and the rank of $R'M_iR$ equals the rank of E_i. But then $(F_i, G_i) = T_i(E_i, F_i')$ where T_i is an $(n - p) \times p$ matrix and $G_i = T_iE_iT_i'$. Let t_i be the first row of T_i. Then the first diagonal element of G_i is $t_iE_it_i'$ which is a sum of squares of some of the elements of t_i and the first row of F_i is t_iE_i which is a vector containing those same elements of t_i and zeros. Hence $\sum G_i = \varphi$ implies that $t_iE_it_i' = 0$ and $t_iE_i = \varphi$. The first row of G_i is $t_iE_iT_i' = \varphi$. Applying this argument to each row of T_i we have $F_i = \varphi$ and $G_i = \varphi$.

Hence

$$R'M_iR = \begin{pmatrix} E_i & \varphi \\ \varphi & \varphi \end{pmatrix}$$

and $R'M_iRR'M_jR = \varphi$ (for all $i \neq j$) and $R'M_iR$ is idempotent. Hence, A_i is idempotent and $A_iA_j = \varphi$ (for all $i \neq j$), and the proof is complete.

It has been pointed out (Craig, [2]) that if Y is distributed as $N_n(\varphi, I)$, then a necessary and sufficient condition that $Y'AY$ be distributed as $\chi^2(p)$ is that A be an idempotent matrix of rank p. We will generalize this result into

THEOREM 2. *If Y is distributed as $N_n(\mu, I)$, then a necessary and sufficient condition that $Y'AY$ is distributed as $\chi'^2(k, \lambda)$ (where $\lambda = \frac{1}{2}\mu'A\mu$) is that A be an idempotent matrix of rank k.*

Proof. We will first prove sufficiency. Let P be an orthogonal matrix such that $P'AP = \begin{pmatrix} I_k & \varphi \\ \varphi & \varphi \end{pmatrix}$, and let $Z = P'Y$. Then $Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$ is distributed as $N_n(\alpha, I)$ where $\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = P'\mu$ and where α_1 and Z_1 are each $k \times 1$ vectors. Z_1 is distributed as $N_k(\alpha_1, I)$. Also $Y'AY = Z'P'APZ = Z_1'Z_1$. Thus by Theorem F, $Y'AY = Z_1'Z_1$ is distributed as $\chi'^2(k, \lambda)$ where $\lambda = \frac{1}{2}\alpha_1'\alpha_1$. This proves sufficiency if we can show that $\alpha_1'\alpha_1 = \mu'A\mu$. To do this let $P = (P_1P_2)$ where P_1 has dimension $n \times k$, then

$$\mu'A\mu = \mu'PP'APP'\mu = \mu'(P_1P_2)P'AP\begin{pmatrix} P_1' \\ P_2' \end{pmatrix}\mu = (\mu'P_1, \mu'P_2)\begin{pmatrix} I_k & \varphi \\ \varphi & \varphi \end{pmatrix}\begin{pmatrix} P_1'\mu \\ P_2'\mu \end{pmatrix}$$

$$= \mu'P_1P_1'\mu = \alpha_1'\alpha_1.$$

To prove necessity, we will assume that $Y'AY$ is distributed as $\chi'^2(k, \lambda)$ and show that this implies that A is an idempotent matrix of rank k. We know that there exists an orthogonal matrix C such that $C'AC = D$ where D is a diagonal matrix where the number of non-zero diagonal elements, d_{ii}, equals the rank of A. Let $Z = C'Y$, then $Y'AY = Z'C'ACZ = Z'DZ = \sum_{i=1}^n d_{ii}z_i^2$. Since Z is distributed as $N_n(C'\mu, I)$, we know by Theorem F that z_i^2 is distributed as $\chi'^2(1, \lambda_i)$ where $\lambda_i = [E(z_i)]^2/2$. Since the z_i are independent the moment generating function of $\sum_{i=1}^n d_{ii}z_i^2$ is

$$\prod_{i=1}^n (1 - 2t\,d_{ii})^{-1/2}e^{-\lambda_i+\lambda_i(1-2d_{ii}t)^{-1}}.$$

Also, since the hypothesis states that $Y'AY$ is distributed as $\chi'^2(k, \lambda)$ (where $\lambda = \frac{1}{2}\mu'A\mu$) the moment generating function of $Y'AY$ is $(1 - 2t)^{-k/2}e^{-\lambda+\lambda(1-2t)^{-1}}$. Since $Y'AY = \sum d_{ii}z_i^2$, the moment generating functions are equal, and we get

$$(2) \qquad (1 - 2t)^{-k/2}e^{-\lambda+\lambda(1-2t)^{-1}} = \prod_{i=1}^n (1 - 2t\,d_{ii})^{-1/2}e^{-\lambda_i+\lambda_i(1-2d_{ii}t)^{-1}}.$$

It is clear that there exists a neighborhood of zero for t such that the quantities on the left- and right-hand sides of Eq. (2) exist and have derivatives of all orders.

If any of the d_{ii} were neither 0 nor 1, the right-hand side of this identity would be an analytic function of t with different singularities than the left-hand side. By this same argument it follows that exactly k of the d_{ii} are one and the others vanish. It also follows that $\lambda = \sum \lambda_i$.

Thus we have shown that if $Y'AY$ is distributed as $\chi'^2(k, \lambda)$, then k of the d_{ii} are equal to unity, and $n - k$ of the d_{ii} are equal to zero. But the d_{ii} are the characteristic roots of A, and hence A is an idempotent matrix of rank k and the theorem is established.

It might be pointed out that $\lambda = 0$, and $\chi'^2(k, \lambda)$ degenerates to $\chi^2(k)$ if and only if $A\mu = \varphi$.

Using Theorem 1 and Theorem 2 of this section and Theorem H of the preceding section, we can state the following Theorems:

THEOREM 3. *If Y is distributed as $N_n(\mu, I)$ and if $Y'AY = \sum_{i=1}^{k} Y'A_iY$ where the rank of A equals p and the rank of A_i equals p_i, then*

(1) *any two of the three conditions* C_1, C_2, C_3 *are necessary and sufficient for all the remaining conditions* $C_1, \cdots, E_1$;

(2) *any two of the three conditions* D_1, D_2, D_3 *are necessary and sufficient for all the remaining conditions,* $C_1, \cdots, E_1$.

(3) *any two conditions* C_i *and* D_j $i \neq j$ *are necessary and sufficient for all the remaining conditions;*

(4) E_1 *and* C_3 *are necessary and sufficient for all the remaining conditions;*

(5) E_1 *and* D_3 *are necessary and sufficient for all the remaining conditions.*

C_1 : $Y'A_iY$ is distributed as $\chi'^2(p_i, \lambda_i)$ where $\lambda_i = (\mu'A_i\mu)/2$ for $i = 1, 2, \cdots, k$.

C_2 : $Y'A_iY$ *and* $Y'A_jY$ *are independent for all* $i \neq j$.

C_3 : $Y'AY$ *is distributed as* $\chi'^2(p, \lambda)$ *where* $\lambda = (\mu'A\mu)/2$.

D_1 : *Each* A_i *is an idempotent matrix.*

D_2 : $A_iA_j = \varphi$ *for all* $i \neq j$.

D_3 : A *is an idempotent matrix.*

E_1 : $\sum_{i=1}^{k} p_i = p$.

THEOREM 4. *In Theorems 2 and 3 if Y is distributed as $N_n(\mu, \sigma^2 I)$ then all the results follow except each quadratic form and each λ and λ_i must be divided by σ^2.*

Cochran's theorem states: *if Y is distributed as $N_n(\varphi, I)$, and if*

$$Y'Y = \sum_{i=1}^{k} Y'A_iY$$

(where the rank of A_i is n_i), then a necessary and sufficient condition that $Y'A_iY$ $(i = 1, 2, \cdots, k)$ are independently distributed respectively as $\chi^2(n_i)(i = 1, 2, \cdots, k)$ is that $\sum_{i=1}^{k} n_i = n$. Madow extended this to (Madow, 1940): *if Y is distributed as $N_n(\mu, I)$ and if $Y'Y = \sum_{i=1}^{k} Y'A_iY$ (where the rank of A_i is n_i), then a necessary and sufficient condition that $Y'A_iY$ $(i = 1, 2, \cdots, k)$ are independently distributed as $\chi'^2(n_i, \lambda_i)$ is that $\sum_{i=1}^{k} n_i = n$.*

We will now extend these theorems.

THEOREM 5. *If Y is distributed as $N_n(\mu, V)$ where V is an $n \times n$ positive definite symmetric matrix, and if $Y'BY = \sum_{i=1}^{k} Y'B_iY$ where the rank of B_i is p_i and the rank of B is p, then any one of the six conditions,* $C_1, C_2, C_3, C_4, C_5, C_6$, *is necessary and sufficient that the $Y'B_iY$ be independently distributed as $\chi'^2(p_i, \lambda_i)$ where $\lambda_i = \frac{1}{2}\mu'A_i\mu$.*

C_1 : BV *be idempotent and* $\sum_{i=1}^{k} p_i = p$.

C_2 : BV *and each* B_iV *be idempotent.*

C_3 : BV *be idempotent and* $B_iVB_j = \varphi$ *for all* $i \neq j$.

C_4 : $Y'BY$ *be distributed as* $\chi'^2(p, \lambda)$ *and* $p = \sum_{i=1}^{k} p_i$. $(\lambda = \frac{1}{2}\mu'A\mu)$.

C_5 : $Y'BY$ *be distributed as* $\chi'^2(p, \lambda)$ *and* B_iV *be idempotent* (where $\lambda = \frac{1}{2}\mu'B\mu$).

C_6 : *$Y'BY$ be distributed as $\chi'^2(p, \lambda)$ and $B_iVB_j = \varphi$ for $i \neq j$* (where $\lambda = \frac{1}{2}\mu'B\mu$).

Proof. Since V is positive definite, there exists a non-singular matrix P such that $P'VP = I_n$. Let $Z = P'Y$; then Z is distributed as $N_n(P'\mu, I_n)$. Also $Y'BY = Z'P^{-1}BP'^{-1}Z$, $Y'B_iY = Z'P^{-1}B_iP'^{-1}Z$, and

$$Z'(P^{-1}BP'^{-1})Z = \sum_{i=1}^{k} Z'(P^{-1}B_iP'^{-1})Z.$$

If we let $A = P^{-1}BP'^{-1}$ and $A_i = P^{-1}B_iP'^{-1}$, then we have $Z'AZ = \sum_{i=1}^{k} Z'A_iZ$, and the results follow immediately from Theorem 3 if we can show that: A being idempotent is equivalent to BV being idempotent; A_i being idempotent is equivalent to B_iV being idempotent; $B_iVB_j = \varphi$ for $i \neq j$ is equivalent to $A_iA_j = 0$ for $i \neq j$. To show these we proceed as follows: If A is idempotent then this means $(P^{-1}BP'^{-1})(P^{-1}BP'^{-1}) = P^{-1}BP'^{-1}$. Performing left multiplication by P and right multiplication by P' gives $BP'^{-1}P^{-1}B = B$. But $P'^{-1}P^{-1} = V$, hence $BVB = B$ or $(BV)(BV) = BV$. Thus A being idempotent implies that BV is idempotent. Starting with $(BV)(BV) = BV$ we will arrive at $AA = A$. Hence A being idempotent is equivalent to BV being idempotent. A similar procedure will work when applied to B_iV. To show that $B_iVB_j = \varphi$ for $i \neq j$ is equivalent to $A_iA_j = \varphi$ for $i \neq j$, proceed as follows: If $B_iVB_j = \varphi$ for $i \neq j$, then $\varphi = P^{-1}B_iP'^{-1}P'VPP^{-1}B_jP'^{-1} = A_iIA_j = A_iA_j$. The reverse procedure also follows, and hence the theorem is established.

(In this theorem, AV and the A_iV need not be symmetric. Also it should be remembered that AV being idempotent is equivalent to VA being idempotent and similarly for A_iV).

We also noted that putting $k = 1$ we get the

COROLLARY 5.1. *If Y is distributed as $N_n(\mu, V)$ where V is a positive definite matrix, then a necessary and sufficient condition that $Y'AY$ be distributed as $\chi'^2(p, \lambda)$ where p is the rank of A and where $\lambda = \frac{1}{2}\mu'A\mu$ is that AV be idempotent (not necessarily symmetric).*

4. Illustrations. Consider the linear hypothesis model $Y = X\beta + e$ defined in Sec. 3. If we partition the X matrix and β vector such that

$$X = (X_1, X_2) \quad \text{and} \quad \beta = \begin{pmatrix} \alpha \\ \gamma \end{pmatrix}$$

where X_1 is of order $n \times p_1$ and α is a $p_1 \times 1$ vector, then we can write $Y = X\beta + e$ as $Y = X_1\alpha + X_2\gamma + e$.

To test the hypothesis H_0 : $\alpha = \varphi$, we can form the ratio

$$u = \frac{Q_1}{Q} \cdot \frac{n - p}{p_1}, \tag{4.1}$$

where u is distributed as Snedecor's F with p_1 and $n - p$ degrees of freedom. The quantities Q_1 and Q can be derived (Kempthorne, 1952) by the following process:

Q is the minimum value of $e'e$ with respect to the parameters in the model $Y = X\beta + e = X_1\alpha + X_2\gamma + e$.

$Q_1 = Q - Q_2$ when Q_2 is the minimum value of $e'e$ with respect to the model $Y = X_2\gamma + e$ (the model restricted by H_0). By a straightforward application of a minimization procedure we see that

$$Q = Y'(I - XS^{-1}X')Y = Y'AY \quad \text{and} \quad Q_2 = Y'(I - X_2S_2^{-1}X_2')Y = Y'BY$$

where $S = X'X$, $S_2 = X_2'X_2$, $I - XS^{-1}X' = A$, and $I - X_2S_2^{-1}X_2' = B$. To find the distribution of Q/σ^2 and Q_1/σ^2 the method sometimes employed (Kempthorne, 1952) is quite a complex procedure of finding the ranks of the corresponding matrices A and B and applying Cochran's theorem. An alternative method using theorems on idempotent matrices to obtain the distribution of u when H_0 is true and when $H_1 : \alpha \neq \varphi$ is true is as follows:

Obviously A and B are each idempotent. Since

$$X'(I - XS^{-1}X') = \varphi, \tag{4.2}$$

it is clear that $X_2'(I - XS^{-1}X') = \varphi$ and $X_1'(I - XS^{-1}X') = \varphi$. Let $C = B - A$; then by using 4.2,

$$C = (I - X_2S_2^{-1}X_2') - (I - XS^{-1}X')$$

is clearly idempotent and $AC = \varphi$. Hence by Theorem 3 we have

1. $Q/\sigma^2 = (Y'AY)/\sigma^2$ is distributed as $\chi'^2(n - p, \lambda_A)$.
2. $Q_1/\sigma^2 = (Y'CY)/\sigma^2$ is distributed as $\chi'^2(p_1, \lambda_C)$.
3. Q and Q_1 are independent.
4. $\lambda_A = 1/2\sigma^2\ (\beta'X'AX\beta) = 1/2\sigma^2\ [\beta'X'(I - XS^{-1}X')X\beta] = 0$, so Q/σ^2 is distributed as $\chi^2(n - p)$.
5. $\lambda_C = [1/(2\sigma^2)]\ [\beta'X'\{(I - X_2S_2^{-1}X_2') - (I - XS^{-1}X')\}X\beta]$
   $= [1/(2\sigma^2)]\ [\alpha'(X_1'X_1 - X_1'X_2S_2^{-1}X_2'X_1)\alpha]$,
   and since $X_1'X_1 - X_1'X_2S_2^{-1}X_2'X_1$ is positive definite, Q_1/σ^2 has the central Chi-square distribution if and only if $\alpha = \varphi$; i.e., if and only if H_0 is true.

Hence by Theorem M, $u = (Q_1/Q)\cdot[(n - p)/p_1]$ is distributed as $F'(p_1, n - p, \lambda_C)$ and reduces to the central F (Snedecor's F) if and only if H_0 is true.

REFERENCES

[1] A. C. Aitken, "On the statistical independence of quadratic forms in normal variates," *Biometrika*, Vol. 37 (1950), pp. 93–96.

[2] A. T. Craig, "Note on the independence of certain quadratic forms," *Ann. Math. Stat.*, Vol. 14 (1943), pp. 195–197.

[3] P. C. Tang, "The power function of the analysis of variance tests with tables and illustrations for their use," *Stat. Res. Memoirs.*, Vol. 2 (1938), pp. 126–149.

[4] H. Hotelling, "On a matric theorem of A. T. Craig," *Ann. Math. Stat.*, Vol. 15 (1944), pp. 427–429.

[5] W. G. Cochran, "The distribution of quadratic forms in a normal system," *Proc. Camb. Phil. Soc.*, Vol. 30 (1934), pp. 178.

[6] W. Madow, "The distribution of quadratic forms in non-central normal random variables," *Ann. Math. Stat.*, Vol. 11 (1940), pp. 100–101.

[7] O. Kempthorne, *The Design and Analysis of Experiments*, John Wiley and Sons, 1952, Chapters 5 and 6.

SOME EXAMPLES WITH FIDUCIAL RELEVANCE[1]

By John W. Tukey

Princeton University

1. Summary. It has been believed by some ([17], p. 204, and perhaps, by implication [21], p. 2, near line 10) that—and R. A. Fisher [e.g. at the Lake Junaluska conference in 1946] has urged the desirability of determining whether—the distribution induced by a pivotal, sufficient and smoothly invertible set of quantities is unique (that is to say, the induced distribution is independent of the choice of a particular set of pivotal quantities among those sets with these properties). If true, such uniqueness would be important in connection with the theory of fiducial probability.

It is the purpose of this paper to present certain examples of particular interest showing that these conditions do *not* provide uniqueness. The first example applies to any family of two-dimensional normal distributions with fixed and known variances and covariances. A one-parameter family of pivotal pairs of quantities are provided, such that no two of the induced distributions are the same. Each pair is sufficient, and consists of two independent quantities, each distributed according to a unit normal distribution. Each pair is shown to be smoothly invertible of every finite order. This example can be extended to the Behrens-Fisher situation. The second example is due to L. J. Savage, and exhibits a two-parameter situation where the two alternative pairs of pivotal quantities constructed according to the prescription of Segal [24] give rise to different distributions.

Mauldon [19] has recently published a quite different example of nonuniqueness which is also based on the bivariate normal distribution. In his example, the means are known and the second moments are to be estimated, so that there are 3 essential parameters.

The paper concludes with a reasonably complete bibliography of papers on fiducial probability.

2. Introduction. The history of fiducial inference has been clouded with dispute and failures of understanding—possibly, however, to no greater extent than is reasonably to be expected when basic new concepts are being forged between the hammer of mathematics and the anvil of concrete applications. This is not the place to review this history, to try to describe fiducial inference as it appears today, or to compare it with other schemes of inference (even as seen by one person). [The writer hopes to do the latter two of these elsewhere (cf. [25]).]

The uniqueness of the result of the fiducial argument has been held by Fisher to be of central importance, and conditions which ensure, or might ensure, unique-

Received April 9, 1956; revised February 1, 1957.

[1] Prepared in connection with research sponsored by the Office of Naval Research.

ness have been important to him and his colleagues (e.g., [21]). The uniqueness problem does not seem to have received the attention which it deserved, even though it was a relatively completely formulated mathematical problem, and could be discussed without touching on any of the relatively sensitive issues of philosophy or principle associated with fiducial inference.

The main example given here appeared first in a paper on "The Purposes of Fiducial Inference" which was presented, as part of a symposium on "Probability and Statistical Inference", to the Econometric Society and the Institute of Mathematical Statistics in Minneapolis on September 6, 1951. The example, as presented at that time, was formally the same, but no detailed proof of smooth invertibility was given, and the need for one was not adequately recognized. The present improvements both in scope and simplicity of approach are the results of comments and stimulation by L. J. Savage, to whom go the author's best thanks.

The example showing that the Segal construction can lead to nonuniqueness is due entirely to Savage, and dates from about the same time (summer, 1952). It was first communicated to the writer in November, 1952.

In so far as examples of nonuniqueness are concerned, the earliest seems to be an unpublished example of P. H. Diananda. Savage informed the writer (in November, 1952) that this is referred to in a Cambridge thesis [29], also unpublished, of R. M. Williams. M. S. Bartlett says that this example is based on the Wishart distribution, and obtains different results in a symmetrical way by starting with the variance of first one variable and then the other. It thus presumably belongs to the same class as Savage's example, and is probably somewhat more complicated to discuss.

More recently, Mauldon has published an explicit example, and has indicated the existence of a family of examples based on the second moments of a family of normal distributions with known, fixed first moments. His examples involve a minimum of 3 parameters, and are thus somewhat more complex. His explicit example involves the permutation of a pair of observations and parameters, as does Savage's example, but in a somewhat less informative situation, since the set of pivotal quantities used were not constructed according to a general method of independent interest, as was the case with Savage's example. The indicated extensions include continuous-parameter families of alternatives, but seem likely to be less understandable than the continuous-parameter cases described here. (Charles Stein has pointed out a way of looking at Mauldon's example in terms of the subgroup of triangular matrices and its conjugate subgroups which makes this example more interesting.)

So far as I know, these are the available examples concerning the nonuniqueness of the result of the fiducial argument (as described in [33]) under varying assumptions. (Conditions under which uniqueness clearly exists can be given, but we shall not discuss them here.)

3. Formalities. A specification determines the probability distribution as a function of parameters. A *quantity* is a function of observations *and* parameters,

defined for all possible combinations of admissible values of its arguments. (Values of the parameters are appropriate as arguments whether or not they determine the distribution of the observations concerned.) A set of quantities is *pivotal* (with respect to a specification) if, whatever admissible values of parameters are inserted in the quantities and are, at the same time, the parameters determining the distribution of the observations appearing in the quantities, the resulting set of statistics has the same distribution. A set of quantities is *sufficient* if, when any arbitrary fixed values of the parameters are inserted as arguments, the resulting set of statistics is sufficient for the parameters determining the distribution of the observations. A set of quantities is *smoothly invertible* (of class α), if, when any possible set of observations is inserted as arguments, the mapping from parameters to quantities:

(1) has the same range for any possible set of observations,
(2) is 1 to 1, and hence has a single-valued inverse, and
(3) this inverse is continuous (and has continuous derivatives of all orders up to α).

If a set of quantities are pivotal and smoothly invertible, then each set of possible observations induces a distribution on parameter space. [A distribution is uniquely associated with the pivotal quantities by their pivotal property. Fixing a set of possible observations fixes a 1 to 1 bicontinuous (i.e., continuous in both directions) relation between quantities and parameters which transfers this distribution to parameter space.]

4. Twisted two-dimensional normals. We now start from a family of two-dimensional normal distributions in which all second-degree moments (about the mean) are fixed, but where all locations in the plane are possible. If we introduce an appropriate coordinate system, the specification becomes the following:

> x and y are normally and independently distributed with averages μ and ν and unit variances.

[We shall abbreviate such statements in the form "x and y are NID(μ, ν; I)."]

We deal with one bivariate observation drawn from some distribution of this family. (This observation may, of course, be the vector of means from a sample of n from a population NID(μ, ν; $\sqrt{n}\, I$).)
Then the quantities

$$w_1 = x - \mu,$$

$$w_2 = y - \nu$$

are immediately seen to be NID(0, 0; I) for any μ and ν, and hence to be pivotal. Fixing μ and ν, the values of w_1 and w_2 determine x and y. The latter are surely sufficient, since they specify the sample (of one) completely. Thus w_1, w_2 are sufficient and pivotal quantities, and since they are obviously smoothly invertible of all orders, we have one smoothly induced distribution.

But the fact that the probability density for (w_1, w_2) is constant on circles

about the origin enables us to modify these quantities without disturbing their distribution.

Let $f(\mu, \nu, r)$ be a sufficiently smooth, but otherwise arbitrary, function of three variables. Put

$$r = \sqrt{w_1^2 + w_2^2} = \sqrt{(x - \mu)^2 + (y - \nu)^2},$$

and

$$w_{1f} = (x - \mu) \cos f + (y - \nu) \sin f = w_1 \cos f + w_2 \sin f,$$

$$w_{2f} = -(x - \mu) \sin f + (y - \nu) \cos f = -w_1 \sin f + w_2 \cos f,$$

where $f = f(\mu, \nu, r)$. Then w_{1f} and w_{2f} are, for each fixed μ and ν, also NID(0, 0; I) and, moreover, we have $r^2 = w_{1f}^2 + w_{2f}^2$. For we have merely twisted each circle of radius r through the angle $f(\mu, \nu, r)$. Each of these pairs (w_{1f}, w_{2f}) is pivotal, and sufficient (since $r^2 = w_{1f}^2 + w_{2f}^2$ is known as soon as w_{1f} and w_{2f} are known, so that fixing μ and ν fixes r and hence make w_1 and w_2 available, and thus makes x and y available!), and, if they are smoothly invertible, are candidates for the construction of our counterexample.

We shall have use for the Jacobian of the transformation from μ, ν to w_{1f}, w_{2f} (with x, y held fixed). The details of calculation are presented in a later section, but the result is

$$J = \frac{\partial(w_{1f}, w_{2f})}{\partial(\mu, \nu)} = 1 + (x - \mu) \frac{\partial f}{\partial \nu} - (y - \nu) \frac{\partial f}{\partial \mu} = 1 + \begin{vmatrix} x - \mu & f_\mu \\ y - \nu & f_\nu \end{vmatrix},$$

where the partial derivatives of f are taken considering f as a function of the three independent variables μ, ν and r. It is shown in the same section that solutions μ, ν of

$$w_{1f}(x, y, \mu, \nu) = a,$$

$$w_{2f}(x, y, \mu, \nu) = b,$$

exist for any chosen x, y, a and b for any continuous $f(\mu, \nu, r)$ and that, if the Jacobian above is always positive, then these solutions are unique, and as many times continuously differentiable as f_μ and J themselves.

5. Detailed example. In order to give a detailed counterexample, we specialize $f(\mu, \nu, r)$ to the form

$$f(\mu, \nu, r) = \frac{\beta\mu}{1 + r^2},$$

when our pivotal quantities become, when written out explicitly,

$$w_{1\alpha} = (x - \mu) \cos \frac{\beta\mu}{1 + (x - \mu)^2 + (y - \nu)^2} + (y - \nu) \sin \frac{\beta\mu}{1 + (x - \mu)^2 + (y - \nu)^2},$$

$$w_{2\alpha} = -(x - \mu) \sin \frac{\beta\mu}{1 + (x - \mu)^2 + (y - \nu)^2} + (y - \nu) \cos \frac{\beta\mu}{1 + (x - \mu)^2 + (y - \nu)^2}.$$

Since $f_\nu = 0$ and $f_\mu = \beta/(1 + r^2)$ we find that

$$\frac{\partial(w_{1\alpha}, w_{2\alpha})}{\partial(\mu, \nu)} = 1 - \beta \frac{y - \nu}{1 + (x - \mu)^2 + (y - \nu)^2},$$

which is surely positive for $|\beta| < 2$. To complete the example, we have only to show that different values of β lead, for one and the same set of observations x, y, to different induced distributions for (μ, ν).

In view of the existence and continuity of all derivatives the induced density in the (μ, ν) plane is

$$\frac{\partial(w_{1\beta}, w_{2\beta})}{\partial(\mu, \nu)} \text{ (density for } w_{1\beta}, w_{2\beta}) = \left(1 - \beta \frac{y - \nu}{1 + r^2}\right) \frac{1}{2\pi} e^{-\frac{1}{2}r^2},$$

where $r^2 = (x - \mu)^2 + (y - \nu)^2$. This induced density is clearly different for different values of β. Subject, then to the verification of (i) the general value of the Jacobian, (ii) the general existence of solutions, and (iii) the uniqueness of solutions when the Jacobian is positive, the announced example is complete.

6. The existence and uniqueness of solutions, and the value of the Jacobian. We now investigate the existence and uniqueness of solutions, μ, ν of

$$w_{1f}(x, y, \mu, \nu) = a,$$
$$w_{2f}(x, y, \mu, \nu) = b,$$

for any given x, y, a and b. We shall use direct methods, noting that smooth invertibility, under the conditions we use is also a consequence of the following result:

Any α times continuously differentiable mapping from a connected open domain to a simply connected range whose Jacobian determinant is continuous and of constant sign, and whose inverse carries compact sets into compact sets, is smoothly invertible of order α. This invertibility theorem is proved elsewhere [26], and examples are given there to show that no one of its topological conditions can be omitted.

Introduce polar coordinates for (a, b) through

$$a = \rho \cos \theta, \qquad b = \rho \sin \theta$$

and polar-like coordinates for μ, ν by

$$\mu = x - r \cos A,$$
$$\nu = y - r \sin A,$$

then the equations to be solved for r and A are, when converted to polar co-

ordinate form, $r = \rho$, and

$$A - f(\mu, \nu, r) = A - f(x - r\cos A, y - r\sin A, r) = \theta + 2k\pi, \ (k \text{ an integer}).$$

Clearly we need only study the second equation

$$\varphi(A) = A - f(x - \rho\cos A, y - \rho(\sin A, \rho) = \theta + 2k\pi \ (k \text{ an integer})$$

for given x, y, ρ, θ and see when it always has a unique solution.

Moreover, we may use these auxiliary variables to evaluate the Jacobian J of w_{1f} and w_{2f}, or, equivalently, of a and b, with respect to μ and ν. We have

$$\begin{aligned}\frac{\partial(a, b)}{\partial(\mu, \nu)} &= \frac{\partial(a, b)}{\partial(\rho, \theta)} \frac{\partial(\rho, \theta)}{\partial(r, A)} \Big/ \frac{\partial(\mu, \nu)}{\partial(r, A)} \\ &= \rho \frac{\partial(\rho, \theta)}{\partial(r, A)} \Big/ \rho = \frac{\partial(\rho, \theta)}{\partial(r, A)} \\ &= \begin{vmatrix} 1 & 0 \\ \dfrac{\partial\theta}{\partial r} & \dfrac{\partial\theta}{\partial A} \end{vmatrix} = \frac{\partial\theta}{\partial r} = \varphi(A)\end{aligned}$$

where we have used $\rho = r$ as necessary. Thus if we evaluate $\varphi(A)$ we will also evaluate the desired Jacobian J.

As A increases from 0 to 2π, the value of $\varphi(v)$ varies from

$$\varphi(0) = -f(x - \rho, y, \rho) \text{ to } \varphi(2\pi) = 2\pi - f(x - \rho, y, \rho) = \varphi(0) + 2\pi$$

and since it is continuous it must pass through the value $\theta + 2k\pi$ for some integer i. Thus we can always solve the initial pair of equations for any $f(\mu, \nu, r)$ which s continuous in its arguments (at least in its first and third arguments together.)

If now we show that $\varphi(A)$ is strictly increasing, it will follow that it takes every value only once, and that solutions not only exist but are unique. Clearly,

$$\begin{aligned}J = \varphi(A) &= 1 - (\rho\sin A)f_\mu + \rho\cos A f_\nu \\ &= 1 - (r\sin A)f_\mu + r\cos A f_\nu \\ &= 1 - (y - \nu)f_\mu + (x - \mu)f_\nu,\end{aligned}$$

and if the Jacobian is always positive, $\varphi(A)$ is monotone, and solutions not only exist but are unique.

Moreover, since

$$\frac{\partial\theta}{\partial A} = \varphi(A) = J, \quad \frac{\partial\theta}{\partial r} = f_r,$$

$$\frac{\partial\rho}{\partial r} = 1, \quad \frac{\partial\rho}{\partial A} = 0,$$

the unique inverse is clearly as many times continuously differentiable as are f_r and the Jacobian.

7. The Behrens-Fisher problem. The best-known example to which the fiducial argument has been applied is the so-called Behrens-Fisher problem, first treated by W. V. Behrens [31] and afterwards discussed by many writers (see Breny [32] for a recent review and [33] for Fisher's most recent statement). The problem arises from two samples, one of n_1 observations from a normal distribution with average μ_1 and the variance $n_1\sigma_1^2$, and other of n_2 observations from a normal distribution with average μ_2 and variance $n_2\sigma_2^2$, when it is desired to make an inference about $\mu_1 - \mu_2$. (For convenience we have defined σ_1 and σ_2 in an unusual way.)

If we take x_1 and x_2 as the means of the two samples, and s_1^2 and s_2^2 as the conventional estimates of the variances of these means, then

$$w_1 = \frac{x_1 - \mu_1}{\sigma_1}, \qquad w_2 = \frac{x_2 - \mu_2}{\sigma_2},$$

$$w_3 = \frac{s_1}{\sigma_1}, \qquad w_4 = \frac{s_2}{\sigma_2},$$

are independently distributed sufficient pivotal quantities. The distribution of w_1 and w_2 is unit bivariate normal, so that w_{1f} and w_{2f} may be introduced as before, and (because μ_1 and μ_2 do not appear in w_3 or w_4, etc.)

$$\frac{\partial(w_{1f}, w_{2f}, w_3, w_4)}{\partial(\mu_1, \mu_2, \sigma_1, \sigma_2)} = \frac{\partial(w_{1f}, w_{2f})}{\partial(\mu_1, \mu_2)} \left(\frac{\partial w_3}{\partial \sigma_1}\right)\left(\frac{\partial w_4}{\partial \sigma_2}\right)$$

and if the Jacobian of w_{1f}, w_{2f}, with respect to μ_1, μ_2 depends on α (as it does) so too does the Jacobian of all four pivotals with respect to all four parameters. Consequently the induced distribution on parameter space is not unique for the Behrens-Fisher specification. (It might be interesting to examine other distributions than the classical one.)

8. Savage's Example. We now set forth the example due to L. J. Savage, which shows how uniqueness can escape us in a different way.

Let x and y be distributed according to

$$\psi(x, y \mid \alpha, \beta)\, dx\, dy = \frac{\alpha^2\beta^2}{\alpha + \beta}(x + y)e^{-(\alpha x + \beta y)}\, dx\, dy,$$

where α and β are positive and $0 \leqq x, y < \infty$. If the cumulative distribution of x is $1 - S$. and the cumulative conditional distribution of y given x is $1 - T$ then

$$S = \left(1 + \frac{\alpha\beta x}{\alpha + \beta}\right)e^{-\alpha x},$$

$$T = \left(1 + \frac{\beta y}{1 + \beta x}\right)e^{-\beta y},$$

as is easily verified by integration. From their definition these quantities are uniformly distributed on $0 \leqq S, T \leqq 1$ and are pivotal. (They are essentially examples of the pivotal quantities pointed out by Segal [24].) Moreover, for x and y fixed, T takes any values between 0 and 1 for suitably chosen β—and for

x, y and β fixed, S takes all values between 0 and 1 for suitably chosen α—thus the open square is covered for each choice of (x, y).

Now let us calculate the Jacobian from (α, β) to (S, T). Clearly $\partial T/\partial\alpha$ vanishes, so that $\partial S/\partial\beta$ is not involved. The other two derivatives are

$$\frac{\partial S}{\partial \alpha} = -\frac{\alpha x}{(\alpha+\beta)^2}[(\alpha+2\beta)+\beta(\alpha+\beta)x]e^{-\alpha x},$$

$$\frac{\partial T}{\partial \beta} = -\frac{y}{(1+\beta)^2}[(1+\beta x)(1+\beta x+\beta y)-1]e^{-\beta y}$$

and are each clearly negative for all positive α, β, x, y. The Jacobian is their product and is clearly positive.

The induced density on $0 \leqq \alpha, \beta < +\infty$ is given by the pivotal density (identically unity) multiplied by the Jacobian, namely

$$\frac{\alpha x y}{(\alpha+\beta)^2} e^{-\alpha x-\beta y} \frac{[\alpha+2\beta+\beta(\alpha+\beta)x][(1+\beta x)(1+\beta x+\beta y)-1]}{(1+\beta x)^2}\, d\alpha\, d\beta,$$

where we have separated a factor symmetric in (α, x) and (β, y) from a factor manifestly not so symmetric.

The specification at the beginning of this section was symmetric in (α, x) and (β, y) so that if we take $1 - S$ as the cumulative distribution of y, and $1 - T$ as the cumulative conditional distribution of x given y and go through the identical argument, we will find that the induced distribution for α, β is obtained from the above by symmetry—by interchanging α with β and x with y. The result will clearly not be the same!

Thus the two applications of Segal's process to this specification do not lead to the same induced distribution. As a consequence we see that conditions of monotony of pivotal quantities, though they enforce smooth invertibility, cannot enforce uniqueness of induced distribution. For S and T are monotone in all of x, y, α, β, as follows when we note

$$\frac{\partial S}{\partial \beta} = \frac{\alpha^2 x}{(\alpha+\beta)^2} e^{-\alpha x} > 0$$

and use our earlier results.

REFERENCES

[1] M. S. Bartlett, "The information available in small samples," *Proc. Cambridge Philos. Soc.*, Vol. 32 (1936), pp. 560–566.

[2] M. S. Bartlett, "Complete simultaneous fiducial distributions," *Ann. Math. Stat.*, Vol. 10 (1939), pp. 129–137.

[3] M. S. Bartlett, "Interpretation of quasi sufficiency," *Biometrika*, Vol. 31 (1940), pp. 391–392.

[4] M. A. Creasy, "Limits for the ratio of means" *J. Roy. Stat. Soc.*, Series B, Vol. 16 (1954), pp. 186–194 (and discussion, pp. 204–222).

[5] E. C. Fieller, "Some problems in interval estimation," *J. Roy. Stat. Soc.*, Series B, Vol. 16 (1954), pp. 175–185 (and discussion, pp. 204–222).

[6] R. A. Fisher, "Inverse probability," *Proc. Cambridge Philos. Soc.* Vol. 26 (1930), pp. 528–535 (reprinted in [16] as paper 22).

[7] R. A. Fisher, "The concepts of inverse probability and fiducial probability referring to unknown parameters," *Proc. Roy. Soc.* (London), Vol. A139 (1933), pp. 343–348.

[8] R. A. Fisher, "Two new properties of mathematical likelihood," *Proc. Roy. Soc.* (London), Vol. A144 (1934), pp. 285–307 (reprinted in [16] as paper 24).
[9] R. A. Fisher, "The logic of uncertain inference," *J. Roy. Stat. Soc.*, Vol. 98 (1935), pp. 39–54 (reprinted in [16] as paper 26).
[10] R. A. Fisher, "The fiducial argument in statistical inference," *Ann. Eugenics*, Vol. 6 (1935), pp. 391–398 (reprinted in [16] as paper 25).
[11] R. A. Fisher, "Uncertain inference," *Proc. Amer. Acad. Arts Sci.*, Vol. 71 (1936), pp. 245–258 (reprinted in [16] as paper 27.)
[12] R. A. Fisher, "On a point raised by M. S. Bartlett on fiducial probability," *Ann. Eugenics*, Vol. 7 (1937), pp. 370–375.
[13] R. A. Fisher, "The comparison of samples with possibly unequal variances," *Ann. Eugenics*, Vol. 9 (1939), pp. 174–180 (reprinted in [16] as paper 35).
[14] R. A. Fisher, "A note on fiducial inference," *Ann. Math. Stat.*, Vol. 10 (1939), pp. 383–388.
[15] R. A. Fisher, "Conclusions fiduciares," *Ann. Inst. Henri Poincaré*, Vol. 10 (1948), pp. 191–213.
[16] R. A. Fisher, *Contributions to Mathematical Statistics*, John Wiley and Sons, New York, 1950.
[17] J. O. Irwin, "Discussion" (of [4] and [5]), *J. Roy. Stat. Soc.*, Series B, Vol. 16 (1954), pp. 204–206.
[18] M. G. Kendall, "On the reconciliation of theories of probability," *Biometrika*, Vol. 36 (1949), pp. 101–116.
[19] J. G. Mauldon, "Pivotal quantities for Wishart's and related distributions, and a paradox in fiducial theory," *J. Roy. Stat. Soc.*, Series B, Vol. 17 (1955), pp. 79–85.
[20] J. Neyman, "Fiducial argument and the theory of confidence intervals," *Biometrika*, Vol. 32 (1941–42), pp. 128–150.
[21] A. R. G. Owen, "Ancilliary statistics and fiducial distributions," *Sankhya*, Vol. 9 (1948), pp. 1–18.
[22] E. S. Pearson, "Note on Professor Pitman's contribution to the theory of estimation," *Biometrika*, Vol. 30 (1938–39), pp. 471–474.
[23] E. J. G. Pitman, "The estimation of the location and scale parameters of a continuous population of any given form," *Biometrika*, Vol. 30 (1938–39), pp. 391–421.
[24] I. E. Segal, "Fiducial distribution of several parameters with application to a normal system," *Proc. Cambridge Philos. Soc.*, Vol. 34 (1938), pp. 41–47.
[25] J. W. Tukey, "The present state of fiducial probability," in preparation.
[26] J. W. Tukey, "A smooth invertibility theorem," to appear, these *Annals*.
[27] B. L. Welch, "On confidence limits and sufficiency, with particular reference to parameters of location," *Ann. Math. Stat.*, Vol. 10 (1939), pp. 58–69.
[28] S. S. Wilks, "Fiducial distributions in fiducial inference," *Ann. Math. Stat.*, Vol. 9 (1938), pp. 272–280.
[29] R. M. Williams, "The use of fiducial distributions with special reference to the Behrens-Fisher problem," Part II of an unpublished dissertation submitted to the University of Cambridge and filed (1949) in its Library as Ph.D. 1671.
[30] Frank Yates, "An apparent inconsistency arising from tests of significance based on fiducial distributions of unknown parameters," *Proc. Cambridge Philos. Soc.*, Vol. 35 (1939), pp. 579–591.
[31] W. V. Behrens, "Ein Betrag zur Fehlerberechnung bei wenig Beobachtungen," *Landw. Jb.*, Vol. 68 (1929), pp. 807–837.
[32] H. Breny, "L'état actuel du problème de Behrens-Fisher," *Trabajos Estadist.*, Vol. 6 (1955), pp. 111–131.
[33] R. A. Fisher, *Statistical Methods and Scientific Inference*, Edinburgh and London, Oliver and Boyd, 1956, viii + 175 pages.

STATISTICAL PROPERTIES OF INVERSE GAUSSIAN DISTRIBUTIONS. II.

By M. C. K. Tweedie

Virginia Polytechnic Institute

0. Summary. Given a fixed number n of observations on a variate x which has the Inverse Gaussian probability density function

$$\exp\left\{-\frac{\phi^2 x}{2\lambda} + \phi - \frac{\lambda}{2x}\right\}\sqrt{\frac{\lambda}{2\pi x^3}}, \qquad 0 < x < \infty,$$

for which $E(x) = \lambda/\phi = \mu$, it is shown how to find functions of the sample mean m whose expectations can be expressed suitably in terms of the parameter ϕ (or μ). In particular, it is shown that the conditional expectation of any unbiased estimator $\tilde{\kappa}_r$ of the rth cumulant κ_r is

$$E(\tilde{\kappa}_r \mid m) = 2m(\tfrac{1}{2}\lambda n^2)^{r-1}e^{\frac{1}{2}g}\int_1^\infty (u-1)^{2r-3}e^{-\frac{1}{2}gu^2}\,du/(r-2)!$$

where $g = \lambda n/m$. This expectation may be evaluated either by series given in the paper or by using published tables of numerical values of certain functions to which it can be related. The conditional variance of the usual mean square estimator s^2 of κ_2 is also found. These results give an asymptotic series for the conditional variance of a generalization $\chi_s^2 = (n-1)s^2/E(s^2 \mid m)$ of a statistic discussed by Cochran. Exact formulae for the expectation of the statistic s^2/m^2 and its mean square error as an estimator of λ^{-1} are given or described. This statistic is a consistent estimator of λ^{-1} and has asymptotically an efficiency of $\phi/(\phi+3)$.

1. Introduction. An earlier paper [1], which will be called "Paper I," was mainly devoted to the characteristics of an Inverse Gaussian variate and its reciprocal and to the maximum likelihood estimation of the parameters. The work reported in that paper was started some considerable number of years ago because of some unusual features of some experimental data which had been obtained in the physics research laboratories of the University of Reading, England. At a casual examination these data showed a strong tendency for the dispersions to increase when samples were considered with increasing means. This tendency was confirmed by calculating arithmetic means and the sums of the squared deviations from the means for a large number of samples, plotting the two against one another and fitting regression curves, using some rather arbitrary assumptions about weighting. A comparable technique was used by Fisher, Thornton, and Mackenzie [2] in analysing some bacterial counts. Our case led to the question of the effect of Brownian motion in the experiments, and to some unsolved problems in theoretical statistics. Although the values

Received April 9, 1956.

of the parameters in the Reading data were such as to make theoretically precise solutions to these problems unnecessary, some exact and asymptotic results have recently been obtained on the conditional distributions of certain statistics at fixed values of the sample mean, and on the overall distributions of some statistics which can be based on them, on the assumption that the random variation was of the Inverse Gaussian nature which would be caused by Brownian motion. These results are of some theoretical interest even if the physical origin of the basic Inverse Gaussian family of distributions is ignored. It may also be added that in the originating physical experiments there were some relatively minor disturbing factors [3] which made the Brownian motion theory an incomplete statistical model.

2. Some general formulae. In this paper the standard form adopted for the probability density function of an Inverse Gaussian variate x will be

$$f(x; \phi, \lambda) = \exp\left\{-\frac{\phi^2 x}{2\lambda} + \phi - \frac{\lambda}{2x}\right\}\sqrt{\frac{\lambda}{2\pi x^3}}, \tag{1}$$

with $0 < x < \infty$ (cf. (1d) of Paper I). The population mean of x is $\mu = \phi/\lambda$. The integral over $(0, \infty)$ of (1), with complex values for ϕ and λ, is equal to unity if the real parts of ϕ^2/λ and λ are positive. The distribution (1) will rarely appear explicitly; for we shall be mainly concerned with the sample mean and conditional distributions referred to fixed values of the sample mean.

As has previously been shown [4], a Laplacian form for a probability density function facilitates the derivation of the regression mean, against the sample mean, of any statistic which is algebraically independent of the population mean μ but whose overall expectation is a suitable known function of μ. The Inverse Gaussian family is of this form and this property has already been used in Paper I. We now proceed to give some general formulae which will be used subsequently to derive some further statistical results.

We shall use m to stand for the arithmetic mean of a fixed number n of independent observations on (1). As was shown in Paper I, the distribution of m is of the same form as (1), with ϕ replaced by ϕn. In the problems considered in this paper, many of the mathematical operations are most conveniently expressed in terms of the variables defined by

$$g = \lambda n/m, \qquad \theta = \phi n. \tag{2}$$

The probability density function of g is

$$\exp\left(-\tfrac{1}{2}\theta^2 g^{-1} + \theta - \tfrac{1}{2}g\right)/\sqrt{2\pi g} = gf(g; \theta, 1). \tag{3}$$

This is of the same form as the density function of y, the reciprocal of the Inverse Gaussian variate x, as given in (30) in Paper I, with $\lambda = 1$ and $\theta = 1/\mu$.

The moments of g are directly obtainable from (12) and (33) of Paper I. In particular, Var $(g) = \theta + 2$ (cf. (35) of that paper). For convenience of reference, we give

$$
\begin{aligned}
&\theta E(g^{-1}) = 1\\
&\theta^3 E(g^{-2}) = E(g) = \theta + 1\\
&\theta^5 E(g^{-3}) = E(g^2) = \theta^2 + 3\theta + 3\\
&\theta^7 E(g^{-4}) = E(g^3) = \theta^3 + 6\theta^2 + 15\theta + 15\\
&\theta^9 E(g^{-5}) = E(g^4) = \theta^4 + 10\theta^3 + 45\theta^2 + 105\theta + 105\\
&\theta^{11} E(g^{-6}) = E(g^5) = \theta^5 + 15\theta^4 + 105\theta^3 + 420\theta^2 + 945\theta + 945\\
&\theta^{13} E(g^{-7}) = E(g^6) = \theta^6 + 21\theta^5 + 210\theta^4 + 1260\theta^3 + 4725\theta^2\\
&\qquad\qquad + 10395\theta + 10395
\end{aligned} \tag{4}
$$

By simple algebraic operations on these results one can easily find polynomials in g whose expectations are the positive integral powers of θ from the first to the sixth.

Under fairly general conditions on the arbitrary functions $h(u)$ and $l(g)$ which appear in the following equation, it is true that

$$
E\left\{l(g)e^{\frac{1}{2}g}\int_1^\infty h(u)e^{\frac{1}{2}gu^2}\,du \mid \theta\right\} = e^{\theta}\int_1^\infty e^{-\theta u}u^{-1}h(u)E\{l(Gu^{-2}) \mid \theta u\}\,du, \tag{5}
$$

where G is a random variable with $Gf(G; Ou, 1)$ as its probability density function. It is both necessary and sufficient that the integrals and expectations in (5) exist for all the required values of the variables on which they depend. A proof may be established by expressing the expectation on the left as an integral and reversing the order of the integrations with respect to g and u.

To develop the first application of (5), take $l(g) = g^{-1}$. Then, since

$$
E\{G^{-1}u^2 \mid \theta u\} = \theta^{-1}u,
$$

therefore

$$
E\left\{g^{-1}e^{\frac{1}{2}g}\int_1^\infty h(u)e^{-\frac{1}{2}gu^2}\,du \mid \theta\right\} = \theta^{-1}\int_0^\infty e^{-\theta v}h(v+1)\,dv. \tag{6}
$$

From the effective uniqueness of the inverse of the Laplace transform, the variate whose overall expectation is taken on the left of (6) must be equal to the conditional expectation, with a fixed value of g or m, of a variate whose overall expectation can be equated to the expression on the right of (6). The problem of finding the conditional expectation, or regression function, thus becomes the relatively simple one of expressing the overall expectation in terms of θ and then inverting the Laplace transform of $h(v+1)$ which appears on the right of (6). In this procedure for inversion it is permissible to take θ to be complex, if necessary, so long as its real part is positive.

If $\tilde{T}$ is some unbiased estimator of a function $T(\theta)$ of θ of known form, it therefore follows that the conditional expectation of this estimator is

$$
\begin{aligned}
E(\tilde{T} \mid g) &= g^{-1}e^{\frac{1}{2}g}\int_1^\infty h(u)e^{\frac{1}{2}-gu^2}\,du = g^{-1}\int_0^\infty h(v+1)e^{-g(v+\frac{1}{2}v^2)}\,du\\
&= g^{-1}\sum_{r=0}^{\infty}\frac{(-\frac{1}{2}g)^r}{r!}\frac{\partial^{2r}}{\partial g^{2r}}[gT(g)].
\end{aligned} \tag{7}
$$

However, this symbolic result (7) is not always the most convenient one to use, as it sometimes leads to infinite series which can be avoided by introducing functions for which tables of numerical values have been published. When

$$T(\theta) = C\theta^{-s} \tag{8}$$

C and s being independent of θ, the inversion of the transform gives, in an obvious notation,

$$E_g^{-1}(C\theta^{-s}) = E(\bar{T} \mid g) = Cg^{-1}e^{\frac{1}{2}g} \int_1^\infty (u-1)^{s-2}e^{-\frac{1}{2}gu^2}\, du/(s-2)! \tag{9}$$

$$= Cg^{-\frac{1}{2}(s+1)}e^{\frac{1}{2}g}Hh_{s-2}(g^{1/2}), \tag{10}$$

where $Hh_{s-2}(\cdot)$ represents the Hermite polynomial function as given by Jeffreys and Jeffreys ([5], 23.081, who give also convergent and asymptotic series expansions of it. When s is an integer greater than 2, these expansions of the Hermite polynomials are infinite series. However, the right-hand side of (9) can be expressed in closed form in terms of more extensively tabulated functions by integrating it by parts. A convenient general formula, to avoid some rather repetitious work if the requisite Hermite polynomials are not known in this form, is

$$\int_a^\infty f(u)e^{-u^2/2b}\, du = (1-bDP)^{-1}f(u) \mid_{u=0} \int_a^\infty e^{-u^2/2b}\, du + bP(1-bDP)^{-1}f(u) \mid_{u=a} e^{-a^2/2b}. \tag{11}$$

Here a and b are positive constants, $f(u)$ is a polynomial function of u, P is an operator with the general property typified by $PF(u) = [F(u) - F(0)]/u$, $F(u)$ being a polynomial in u, and D is the operator of differentiation with respect to u. The compound symbol $(1 - bDP)^{-1}$ is an abbreviation for the series $1 + bDP(1 + bDP(1 + bDP(1 + \cdots)))$, which $= 1 + bDP + b^2DPDP + b^3DPDPDP + \cdots$.

In (9) the general formula (11) would take $f(u) = (u-1)^{s-2}$, $a = 1$, $b = g^{-1}$. As an example, suppose that $T(\theta) = \theta^{-6}$. Then we shall have

$$\begin{aligned}
f(u) &= u^4 - 4u^3 + 6u^2 - 4u + 1, && \text{which} = \ \ 1 \text{ at } u = 0\\
Pf(u) &= u^3 - 4u^2 + 6u - 4 && \text{, which} = -1 \text{ at } u = 1\\
DPf(u) &= 3u^2 - 8u + 6 && \text{, which} = \ \ 6 \text{ at } u = 0\\
PDPf(u) &= 3u - 8 && \text{, which} = -5 \text{ at } u = 1\\
DPDPf(u) &= 3 && \text{, which} = \ \ 3 \text{ at } u = 0
\end{aligned}$$

Thus

$$E_g^{-1}(\theta^{-6}) = g^{-1}e^{\frac{1}{2}g} \int_1^\infty (u-1)^4e^{-\frac{1}{2}gu^2}\, du/4! \tag{12}$$

$$= g^{-1}\{(1 + 6g^{-1} + 3g^{-2})I - g^{-1} - 5g^{-2}\}/24,$$

where

$$I = e^{\frac{1}{2}\theta}\int_1^\infty e^{-\frac{1}{2}\theta u^2}\,du = e^{-\frac{1}{2}\theta}\sqrt{2\pi/\theta}\int_{\sqrt{\theta}}^\infty e^{-\frac{1}{2}v^2}\,dv/\sqrt{2\pi}, \tag{13}$$

for which Laplace found a continued fraction expansion (cf. [6], p. 263; [7], p. v; [8], Eq. (92.11)). It is of interest that the expression within the braces in (12) is the difference between the numerator of one of the convergents of this continued fraction and I times the denominator of the same convergent, so that a rather large number of significant figures is needed in I if (12) is to be evaluated accurately. A comparable situation occurs with other formulae in this field of study.

The following special formulae, which are obtainable by combining (4) and (5) with $h(u) = 1$, are also useful:

$$E(I) = e^\theta\int_\theta^\infty x^{-1}e^{-x}\,dx, \tag{14}$$

which has a well-known asymptotic series expansion, obtainable by integrating by parts, and may be expressed as a continued fraction (cf. [8], Eq. (92.16)); and

$$\begin{aligned}
2\cdot 1!E(gI) &= -\theta^2E(I) + \theta + 1\\
2^2\cdot 2!E(g^2I) &= \theta^4E(I) - \theta^3 + \theta^2 + 6\theta + 6\\
2^3\cdot 3!E(g^3I) &= -\theta^6E(I) + \theta^5 - \theta^4 + 2\theta^3 + 42\theta^2 + 120\theta + 120\\
2^4\cdot 4!E(g^4I) &= \theta^8E(I) - \theta^7 + \theta^6 - 2\theta^5 + 6\theta^4 + 360\theta^3 + 2040\theta^2\\
&\quad + 5040\theta + 5040\\
2^5\cdot 5!E(g^5I) &= -\theta^{10}E(I) + 0!\theta^9 - 1!\theta^8 + 2!\theta^7 - 3!\theta^6 + 4!\theta^5\\
&\quad + 372\theta^4 + 3528\theta^3 + 15624\theta^2 + 9!\theta + 9!
\end{aligned} \tag{15}$$

In deriving these last formulae (15), it was found convenient to use, on the right side of (5), the identity

$$\begin{aligned}
e^\theta\int_\theta^\infty \sum_{i=0}^{r}(i!c_ix^{-i})x^{-1}e^{-x}\,dx = e^\theta\int_\theta^\infty x^{-1}e^{-x}\,dx(c_0 - c_1 + c_2 - c_3 + \cdots)\\
+ 0!\theta^{-1}(c_1 - c_2 + c_3 - \cdots)\\
+ 1!\theta^{-2}(c_2 - c_3 + \cdots)\\
+ 2!\theta^{-3}(c_3 - \cdots)\\
+ \cdots + (r-1)!\theta^{-r}c_r.
\end{aligned} \tag{16}$$

For completeness, the following results are also recorded:

$$\begin{aligned}
E(g^{-1}I) &= \theta^{-2}\\
E(g^{-2}I) &= \theta^{-3} + 2\theta^{-4}\\
E(g^{-3}I) &= \theta^{-4} + 5\theta^{-5} + 8\theta^{-6}\\
E(g^{-4}I) &= \theta^{-5} + 9\theta^{-6} + 33\theta^{-7} + 48\theta^{-8}.
\end{aligned} \tag{17}$$

In conjunction with (4), these results enable polynomials in g^{-1} to be found which, when multiplied by I, give expressions whose expectations are the negative integral powers of θ from the first to the seventh. The results for the first, third, fifth, sixth and seventh negative powers are given essentially in (12) and (23).

It is also convenient to introduce a further new variable to simplify the presentation of the results in the following sections. We write

$$J = 1 - gI \tag{18a}$$

$$= 1 - ge^{\frac{1}{2}g}\int_1^\infty e^{-\frac{1}{2}gu^2}\,du \tag{18b}$$

$$= g^{-1} - 3g^{-2} + \cdots + (-1)^{r+1}3\cdot5\cdots(2r-1)g^{-r} + (-1)^r 3\cdot5\cdots(2r-1)(2r+1)g^{-r}e^{\frac{1}{2}g}\int_1^\infty u^{-2r-2}e^{-\frac{1}{2}gu^2}\,du. \tag{19}$$

It can be shown that J decreases monotonely from 1 to 0, and that gJ increases monotonely from 0 to 1, when g increases from 0 to infinity.

The form (19) gives an asymptotic series which is satisfactory when g is sufficiently large. A useful expansion for moderately large values of g is the continued fraction

$$J = \cfrac{g^{-1}}{1+\cfrac{3g^{-1}}{1+\cfrac{2g^{-1}}{1+\cfrac{5g^{-1}}{1+\cfrac{4g^{-1}}{1+\text{etc.}}}}}} = \cfrac{1}{g+\cfrac{3}{1+\cfrac{2}{g+\cfrac{5}{1+\cfrac{4}{g+\text{etc.}}}}}} \tag{20}$$

The constants in the successive partial numerators are the positive integers, following the sequence with alternating reversals 1; 3, 2; 5, 4; 7, 6; 9, 8; etc. Alternatively Laplace's continued fraction could be used to find I. When g approaches zero,

$$J = 1 - (\tfrac{1}{2}\pi g)^{1/2} + 0(g). \tag{21}$$

In the notation adopted in the National Bureau of Standards' Tables of Probability Functions ([9], p. xix),

$$I = g^{-1/2}F(g^{1/2}), \qquad J = 1 - L(2^{-1/2}g^{1/2}).$$

Hence a published table of either $F(x)$ or $L(x)$ may be used to shorten the computations, provided it gives enough significant digits. Table I gives I, J and gJ for certain values of g, chosen because Sheppard's 1939 tables [7] could be used for them without interpolation. Burgess's Table 3 [6], which is unfortunately 'seriously infested with error' according to page x of Sheppard's tables [7], would similarly lead directly to values for which $\frac{1}{2}g$ is an exact square.

TABLE I

g	I	J	gJ
1	.65567 95424	.34432 04576	.34432 04576
4	.21068 46146	.15726 15414	.62904 61657
9	.10153 00996	.08622 91039	.77606 19348
16	.05916 30957	.05339 04683	.85424 74935+
25	.03856 16209	.03595 94764	.89898 69106
36	.02706 29435−	.02573 40346	.92642 52463
49	.02001 48834	.01927 07158	.94426 50756
64	.01539 14954	.01494 42939	.95643 48119
81	.01219 85870	.01191 44568	.96507 10004
100	.00990 28596	.00971 40353	.97140 35283

3. Conditional means and variances of certain statistics. The cumulants of the general Inverse Gaussian variate were given in (9) of Paper I. Since they are of the form of (8), we can use (9) of the present paper and get

$$(22)\qquad E(\bar{\kappa}_r \mid m;\lambda,n) = 2m(\tfrac{1}{2}\lambda n^2)^{r-1}e^{\frac{1}{2}g}\int_1^\infty (u-1)^{2r-3}e^{-\frac{1}{2}gu^2}\,du/(r-2)!,$$

from which

$$E_m^{-1}(\kappa_2) = E(\bar{\kappa}_2 \mid m;\lambda,n) = nm^2J = \lambda^{-1}m^3gJ = \lambda^2n^3g^{-2}J$$

$$(23)\qquad E_m^{-1}(\kappa_3) = E(\bar{\kappa}_3 \mid m;\lambda,n) = \tfrac{1}{2}n^2m^3\{(g+3)J-1\}$$

$$E_m^{-1}(\kappa_4) = E(\bar{\kappa}_4 \mid m;\lambda,n) = \tfrac{1}{2}n^3m^4\{(g^2+10g+15)J-g-7\}.$$

These formulae may be verified by using (4) and (17). The terms within the braces in (23) are the same as occur in the successive convergents of the continued fraction expansion (20), so that, in a similar way as the computation of (12), high precision is needed in J to give accurate numerical results. For example, when $g = 25$, if we took $I = 0.0385616$ or $J = 0.035959$ or $L = 0.964041$, which are correct according to the usual rounding-off rules to the last digit shown, the rounding error (which is 0.21 of the last place shown for I, and 0.48 of the last place shown for J and L) would give a value for $E_m^{-1}(\kappa_4)$ which would be 12 per cent too high if I were used, or 11 per cent too low if J or L were used.

When g is large, the asymptotic expansion of the Hermite polynomial given by Jeffreys and Jeffreys ([5], 23.082) gives

$$(24)\qquad E(\bar{\kappa}_r \mid m;\lambda,n) \sim \frac{4\lambda}{m}\left(\frac{m^2}{2\lambda}\right)^r \sum_{i=0}^{\infty}\frac{(2i+2r-3)!}{i!(r-2)!(-2g)^i}$$

From this, or by using the asymptotic series expansion (19) for J,

$$E(\bar{\kappa}_2 \mid m;\lambda,n) \sim \lambda^{-1}m^3(1-3g^{-1}+15g^{-2}-105g^{-3}+945g^{-4}-\cdots)$$

$$E(\bar{\kappa}_3 \mid m;\lambda,n) \sim 3\lambda^{-2}m^5(1-10g^{-1}+105g^{-2}-1260g^{-3}$$

$$+ 17325g^{-4} - \cdots)$$

(25)
$$E(\bar{\kappa}_4 \mid m; \lambda, n) \sim 15\lambda^{-3}m^7(1 - 21g^{-1} + 378g^{-2} - 6930g^{-3} + 135135g^{-4} - \cdots).$$

4. Conditional variance of Cochran's χ_c^2 statistic. By the same technique of inverting a Laplace transform, as was shown previously ([4], p. 48), the conditional variance of the usual unbiased mean square estimator—viz, $s^2 = \sum(x - m)^2/(n - 1)$—of κ_2 can be found. We first find

$$E(s^4 \mid m; \lambda, n) = \frac{n+1}{n-1} E_m^{-1}(\kappa_2^2) + \frac{1}{n} E_m^{-1}(\kappa_4) \tag{26}$$

for which an exact expression in terms of I or J can be found by using (12) and (23), since $\kappa_2^2 = \theta^{-6}\lambda^4 n^6$. The required conditional variance is obtainable by subtracting the square of $E(s^2 \mid m; \lambda, n)$ from (26). The asymptotic form is

$$\text{Var}\,(s^2 \mid m; \lambda, n) \sim \frac{2m^6}{(n-1)\lambda^3}\left\{1 + \frac{3(n-6)m}{\lambda n} - \frac{6(12n - 47)m^2}{\lambda^2 n^2} + \frac{30(47n - 152)m^3}{\lambda^3 n^3} - \cdots\right\}. \tag{27}$$

For comparison with some of the results already given [4] for the Inverse Poisson (or chi-square type) and the Poisson types of distribution, and also to provide a more direct comparison with the exact chi-square distribution which occurs in the distribution of the maximum likelihood estimator of λ, we may consider a measure of dispersion equivalent to that studied by Cochran [10] for other distributions:

$$\chi_c^2 = (n-1)s^2/E(s^2 \mid m; \lambda, n). \tag{28}$$

Obviously

$$E(\chi_c^2 \mid m; \lambda, n) = n - 1.$$

By using (25) and (27), we get, when g is large,

$$\text{Var}\,(\chi_c^2 \mid m; \lambda, n) \sim 2(n-1)\left\{1 + \frac{3m}{\lambda}\left(1 - \frac{4}{n}\right) + \frac{54m^2}{\lambda^2 n}\left(1 - \frac{59}{18n}\right) + \right\} \cdots. \tag{29}$$

It will be noticed that the absolute values of the leading terms in the series in both (27) and (29) are minimized, as functions of the sample size n, and simultaneously become very insensitive to the precise value of λ, when n is quite small—between 3 and 6 approximately. With the Inverse Poisson distribution the second term in the series corresponding to (29) was found to vanish when $n = 5$. It may therefore be surmised that the standard chi-square

distribution will be a good approximation to that of χ^2_n with samples of about this size, with either the Inverse Gaussian or the Inverse Poisson distribution, assuming that the correct regression formula $E_m^{-1}(\kappa_2)$ is used in (28). An approximate confidence or fiducial interval for λ may be derived from this result.

5. An approximate estimator of λ^{-1}. The statistic s^2m^{-2} might be used as an estimator of λ^{-1}, as an alternative to the estimator of maximum likelihood discussed in Paper I. The conditional mean and variance of s^2m^{-2} are obtainable from the above results, to provide a partial check on its suitability for this purpose. Thus from (23),

$$E(s^2m^{-2} \mid m; \lambda, n) = \lambda^{-1}gJ = \lambda^{-1}(g - g^2I). \tag{30}$$

The series expansions of this formula and of the formula for the conditional variance are obtainable immediately from (25) and (27). Both the expectation and the variance depend to some extent on m, in contrast to the corresponding results which follow from (22) in Paper I, viz,

$$E\left\{\sum_{i=1}^{n} (x_i^{-1} - m^{-1})/(n-1) \mid m; \lambda, n\right\} = 1/\lambda, \tag{31}$$

$$\mathrm{Var}\left\{\sum_{i=1}^{n} (x_i^{-1} - m^{-1})/(n-1) \mid m; \lambda, n\right\} = 2/\lambda^2(n-1). \tag{32}$$

The overall expectation of s^2m^{-2} is, from (4), (15) and (30),

$$E\{s^2m^{-2} \mid \phi, \lambda, n\} = [2 + 2\phi n - (\phi n)^2 + (\phi n)^3 - (\phi n)^4 e^{\phi n} \int_{\phi n}^{\infty} x^{-1}e^{-x}\,dx]/8\lambda \tag{33}$$

$$= \frac{1}{\lambda}\left\{1 - \frac{1}{2^2 \cdot 2!}\left[\frac{4!}{\phi n} - \frac{5!}{(\phi n)^2} + \cdots + \frac{(-1)^{r-1}(r-1)!}{(\phi n)^{r-4}} + (\phi n)^4 e^{\phi n} \int_{\phi n}^{\infty} \frac{(-1)^r r!}{x^{r+1}} e^{-x}\,dx\right]\right\}. \tag{34}$$

The formula (34) might be used as an asymptotic series for computations, or a continued fraction might be found.

The conditional mean square error, $E(s^4m^{-4} - 2\lambda^{-1}s^2m^{-2} + \lambda^{-2} \mid m; \lambda, n)$, may be evaluated by using the results of (23) and (26). The exact overall mean square error can be found from it by using (4) and (15), but the full expression is too lengthy to be given here. Its asymptotic series expansion is

$$E\{(s^2m^{-2} - \lambda^{-1})^2 \mid \phi, \lambda, n\} \sim \frac{2}{\lambda^2(n-1)}\left\{1 + \frac{3(n-6)}{\phi n} + \frac{15(5n-33)}{(\phi n)^2} + \frac{105(27n-33)}{(\phi n)^3} + \cdots\right\}. \tag{35}$$

This formula is sufficient to show that s^2m^{-2} is a "consistent" estimator of λ^{-1}. By comparison with (31) and (32), its efficiency, on the basis of the mean square error or the variance, is seen to be asymptotically $\phi/(\phi + 3)$.

6. Acknowledgment. The work reported in this paper was done under a grant from the National Science Foundation.

REFERENCES

[1] M. C. K. TWEEDIE, "Statistical properties of Inverse Gaussian distributions. I," *Ann. Math. Stat.*, Vol. 28 (1957), pp. 362–377.

[2] R. A. FISHER, H. G. THORNTON, AND W. A. MACKENZIE, "The accuracy of the plating method of estimating the density of bacterial populations," *Annals of Applied Biology*, Vol. 9 (1922), pp. 325–359.

[3] M. C. K. TWEEDIE, "A mathematical investigation of some electrophoretic measurements on colloids," unpublished thesis for M.Sc. degree, University of Reading, England, 1941.

[4] M. C. K. TWEEDIE, "Functions of a statistical variate with given means, with special reference to Laplacian distributions," *Proc. Cambridge Philos. Society*, Vol. 43 (1947), pp. 41–49.

[5] H. JEFFREYS AND B. S. JEFFREYS, *Methods of Mathematical Physics*, Cambridge University Press, Cambridge, 1946; 3d Ed., 1956.

[6] J. BURGESS, "On the definite integral $2\pi^{-1/2}\int_0^t e^{-t^2}\,dt$, with extended tables of values," *Trans. Roy. Soc. of Edinburgh*, Vol. 39 (1898), pp. 257–321.

[7] W. F. SHEPPARD, *The Probability Integral*, British Association for the Advancement of Science Mathematical Tables, Vol. 7, Cambridge University Press, Cambridge, 1939.

[8] H. S. WALL, *Analytical Theory of Continued Fractions*, Van Nostrand, New York, 1948.

[9] *Tables of Probability Functions*, Vol. 2, National Bureau of Standards, 1942.

[10] W. G. COCHRAN, "The χ^2 distribution for the binomial and Poisson series, with small expectations," *Ann. Eugenics*, London, Vol. 7 (1937), pp. 207–217.

THE MEAN AND VARIANCE OF THE MAXIMUM OF THE ADJUSTED PARTIAL SUMS OF A FINITE NUMBER OF INDEPENDENT NORMAL VARIATES

By M. E. Solari and A. A. Anis

Chelsea Polytechnic, London

1. Introduction. In planning the storage capacity of a reservoir it is desirable to avoid in so far as is practicable both the loss of water that occurs if the reservoir overflows and the harm that is done if the reservoir is empty when water is needed. Hurst [1] on the basis of data from a long series of annual totals of river discharges has discussed the relation between the capacity, the inflow and its variability, and the draft from a reservoir. In the present paper the theoretical analysis of the problem as studied by Anis and Lloyd is carried further.

If, for a period of n years, the annual increment of inflow minus draft is represented by the variable X_i $(i = 1, \cdots, n)$ and the partial sums of these increments by $S_r = \sum_{i=1}^{r} X_i$ $(r = 1, \cdots, n)$, then the maximum U_n over the n-year period of these S_r is the maximum accumulated storage when there is no deficit, their minimum L_n gives the maximum accumulated deficit when there is no storage, and their range $R_n = U_n - L_n$ gives the capacity necessary to avoid the two difficulties mentioned above. Anis and Lloyd [3] have studied the distribution of U_n and R_n for the idealized case in which the X_i are taken as independent standard normal variables and have shown that, for any $n \geqq 2$, the expected value of the maximum is $(2\pi)^{-1/2} \sum_{s=1}^{n-1} s^{-1/2}$ and hence that the asymptotic value of the mean range, which is twice that of the maximum, agrees with the value $2[(2/\pi)n]^{1/2}$ obtained by Feller [2]. Furthermore Anis [4] has shown the second moment about the origin of the maximum to be

$$\frac{n+1}{2} + \frac{1}{2\pi} \sum_{s=2}^{n-1} \sum_{t=1}^{s-1} t^{-\frac{1}{2}}(s-t)^{-\frac{1}{2}}$$

and has obtained [5] a recurrence relation for computing moments of higher order by means of which he has tabulated the values of the first four moments for $n = 2, 3, \cdots, 15$.

However, from both the engineering and the statistical point of view it is sometimes desirable to separate the effect of inflow and draft, since the latter may be controlled in such a way that the former is the decisive random variable. In his paper Hurst considered the effect that would have been obtained by a rule of release which made the annual draft equal to the mean annual inflow for the n-year period, $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$, so that the accumulation after r years became the adjusted partial sum $S'_r = \sum_{i=1}^{r} X_i - r\bar{X}_n$. For these adjusted partial sums Hurst and Feller both obtained $[(\pi/2)n]^{1/2}$ for the asymp-

Received October 18, 1956.

totic mean range.[1] Statistically the study of these adjusted partial sums is advantageous because, since they are distributed about zero provided merely the individual X_i are distributed about a common though not necessarily zero mean, there is now no loss of generality in taking that common mean to be zero.

In this paper we obtain, for the case in which the X_i are independent normal variates with a common mean and unit variance, the distribution of the maximum of the adjusted partial sums and find, for any $n \geqq 2$, the first and second moments about the origin to be

$$\mu_1'(n) = \frac{1}{2}\sqrt{\frac{n}{2\pi}} \sum_{s=1}^{n-1} s^{-\frac{1}{2}}(n-s)^{-\frac{1}{2}}$$

$$\mu_2'(n) = \frac{1}{6}\left\{\frac{n^2-1}{n} + \frac{\sqrt{n}}{2\pi}\sum_{s=2}^{n-1}\sum_{t=1}^{s-1}\frac{s(2s-n)}{\sqrt{(n-s)t^3(s-t)^3}}\right\}$$

with the asymptotic values $\frac{1}{2}[(\pi/2)n]^{1/2}$ and $(n/2) - n^{1/2}$ respectively. Since the distribution of the minimum of the adjusted partial sums is, as in the case of the unadjusted sums, that of minus the maximum, the mean range is twice the mean of the maximum so that our asymptotic value is seen to agree with that obtained by Feller and Hurst.

2. Distribution of the maximum of the adjusted partial sums. In addition to the notation already introduced in Section 1, we shall use throughout $\phi(x)$ to denote the probability density function of a standard normal variate, i.e., $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$. In this connection it should be noted that in accordance with the remarks above, our results will be valid if the annual increments are independent and normally distributed about a common mean with unit variance since reduction to the standard normal variates X_i will not affect the S_r'.

We shall also use $P_n(u)$ and $p_n(u)$ to denote respectively the distribution function and the density function of the maximum over r, U_n', of the adjusted partial sums S_r'. Since by definition S_n' is zero, we consider the $n-1$ sums S_r' $(r = 1, \cdots, n-1)$ and let their maximum be V_n so that $U_n' = \text{Max}\,[V_n, 0]$. Then $P_n(u) = \Pr\{U_n' \leqq u\} = 0$ for $u < 0$ and $P_n(u) = \Pr\{V_n \leqq u\}$ for $u \geqq 0$, so that $P_n(u)$ has a saltus at $u = 0$ and $p_n(u)$ is not defined there. For $u < 0$, $p_n(u) = 0$ and for $u > 0$, $p_n(u) = dP_n(u)/du$.

For any $n \geqq 2$

$$P_n(u) = \int_R (n) \int (2\pi)^{-n/2}\exp(-\tfrac{1}{2}\mathbf{x}'\mathbf{x}) \prod_{i=1}^{n} dx_i \qquad (u \geqq 0), \tag{1}$$

[1] However, the values of the range of these adjusted partial sums observed by Hurst appeared to be more nearly proportional to $n^{.73}$. For this reason the authors of this paper thought the exact formula for the mean range of these adjusted partial sums as a function of n would be of interest.

where the region of integration K is defined by

$$K: \sum_{i=1}^{r} X_i - r\bar{X}_n \leqq u \qquad (r = 1, \cdots, n-1)$$

and $\mathbf{x}$ is the n-dimensional vector of the observations $X_1, \cdots, X_n$. We introduce the transformation

$$\mathbf{x} = \mathbf{B}\mathbf{y}, \tag{2}$$

where $\mathbf{B}$ is the $n \times n$ matrix given by

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 1 \\ -1 & 1 & 0 & \cdots & 1 \\ 0 & -1 & 1 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix}.$$

It is easy to see that the Jacobian J_n of the transformation (2) is given by the recurrence relation

$$J_n = 1 + J_{n-1}$$

and hence that $J_n = n$. Now $\mathbf{x}'\mathbf{x} = \mathbf{y}'\mathbf{B}'\mathbf{B}\,\mathbf{y} = \mathbf{y}'\mathbf{C}\,\mathbf{y}$ where $\mathbf{C}$ is the $n \times n$ matrix given by

$$\mathbf{C} = \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 2 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & n \end{pmatrix}.$$

Hence $\mathbf{x}'\mathbf{x} = \mathbf{y}'\mathbf{C}\,\mathbf{y} = n\,y_n^2 + \mathbf{Y}'\mathbf{A}\,\mathbf{Y}$, where $\mathbf{A}$ is the $(n-1) \times (n-1)$ matrix obtained from $\mathbf{C}$ by omitting the last column and the last row and $\mathbf{Y}$ is the $(n-1)$-dimensional vector $y_1, y_2, \cdots, y_{n-1}$. Reverting to (1) and (2), we see that $y_n = \bar{X}_n$, $y_r = S_r'$ $(r = 1, \cdots, n-1)$ and hence $P_n(u)$ can be put in the form

$$P_n(u) = \sqrt{n} \int_{-\infty}^{u} (n-1) \int_{-\infty}^{u} (2\pi)^{-[(n-1)/2]} \exp\left(-\tfrac{1}{2}\mathbf{Y}'\mathbf{A}\mathbf{Y}\right) \prod_{i=1}^{n-1} dy_i. \tag{3}$$

It is worth noting here that the integrand in (3) is, except for a constant factor, precisely the integrand in expression (6.1) in Anis and Lloyd [3]. Using the value obtained there for that integral, we deduce immediately that

$$P_n(0) = \frac{1}{n}. \tag{4}$$

Differentiating (3) by using the rule for differentiation of multiple integrals when the integrand does not contain the variable with respect to which we

differentiate but the limits of integration do, which is justifiable since our integrand is a well behaved function, we obtain immediately

$$p_n(u) = \sqrt{2\pi n} \sum_{s=0}^{n-2} h_s(u) h_{n-2-s}(u) \qquad (n \geqq 2), \tag{5}$$

where $h_s(u)$ is the integral defined by Anis and Lloyd [3]; i. e., for $s \geq 1$

$$h_s(u) = \int_0^\infty (s) \cdot \int_0^\infty \phi(u - y_1)\phi(y_1 - y_2) \cdots \phi(y_{s-1} - y_s)\phi(y_s)\, dy_1 \cdots dy_s \tag{6}$$

and

$$h_0(u) = \phi(u).$$

Since the probability density functions for the maximums of the unadjusted partial sums are expressible (Anis and Lloyd [3]) as a linear combination of these integrals $h_s(u)$, it is now possible to express $p_n(u)$ in terms of those probability density functions. However, it proves more convenient to obtain the moments of the distribution (5) directly from the properties of the integrals $h_s(u)$.

3. Properties of the integrals $h_s(u)$.

LEMMA 1.

$$h_s(0) = (2\pi)^{-1/2}(s + 1)^{-s/2} \qquad (s \geqq 0). \tag{7}$$

This was proved by Anis and Lloyd [3], since $h_s(0) = (2\pi)^{-[(s+1)/2]} c_s$ in their notation, and is repeated here merely for completeness.

LEMMA 2.

$$h_s(\infty) = 0 \qquad (s \geqq 0). \tag{8}$$

To prove this we note that, by virtue of its definition (6) as an integral, $h_s(u)$ is non-negative for all values of u. Hence the probability density functions $p_n(u)$ are from (5) the sum of non-negative terms $h_s(u)h_{n-s}(u)$ for all n and s. Since $p_n(\infty)$ is zero, no one of these terms and so no $h_s(u)$ can differ from zero at infinity.

LEMMA 3. *For* $s \geqq 1$

$$h_s(u) = \int_0^\infty \phi(u - y)h_{s-1}(y)\, dy, \tag{9}$$

$$h_s'(u) = \int_0^\infty \phi(u - y)yh_{s-1}(y)\, dy - uh_s(u) \tag{10}$$

$$= \int_0^\infty \phi(u - y)h_{s-1}'(y)\, dy + h_{s-1}(0)\phi(u),$$

$$(11)\qquad h_s''(u) = \int_0^\infty \phi(u-y)y^2 h_{s-1}(y)\,dy - 2uh_s'(u) - (u^2+1)h_s(u)$$

$$= \int_0^\infty \phi(u-y)y h_{s-1}'(y)\,dy - uh_s'(u)$$

$$= \int_0^\infty \phi(u-y)h_{s-1}''(y)\,dy + h_{s-1}(0)\phi'(u) + h_{s-1}'(0)\phi(u).$$

To prove this lemma we note that the reduction formula (9) for $h_s(u)$ itself follows immediately from the definition (6) of $h_s(u)$. The reduction formulae (10) and (11) for the derivatives then follow by differentiation of (9) with some rearranging and integration by parts.

LEMMA 4. *For* $s \geqq 1$

$$(12)\qquad h_s(0) = \int_0^\infty h_0(y)h_{s-1}(y)\,dy,$$

$$(13)\qquad h_s'(0) = \int_0^\infty y h_0(y)h_{s-1}(y)\,dy = h_0(0)h_{s-1}(0) + \int_0^\infty h_0(y)h_{s-1}'(y)\,dy,$$

$$(14)\qquad h_s''(0) = \int_0^\infty y^2 h_0(y)h_{s-1}(y)\,dy - h_s(0) = \int_0^\infty y h_0(y)h_{s-1}'(y)\,dy$$

$$= h_0(0)h_{s-1}'(0) + \int_0^\infty h_0(y)h_{s-1}''(y)\,dy.$$

These results follow immediately on putting $u = 0$ in the reduction formulae of Lemma 3.

4. Moments of the distribution of the maximum of the adjusted partial sums. In this paragraph for simplicity of notation we shall omit the limits of integration which will be from zero to infinity throughout and write h_s only wherever we mean this function to be evaluated at zero. Furthermore we shall consider the distribution

$$p_{n+2}(u) = \sqrt{2\pi(n+2)} \sum_{s=0}^{n} h_s(u)h_{n-s}(u) \qquad (n \geqq 0).$$

For this distribution we write the rth moment about the origin

$$\mu_r'(n+2) = \int u^r p_{n+2}(u)\,du \qquad (r \geqq 0)$$

in the form

$$(15)\qquad \mu_r'(n+2) = \sqrt{2\pi(n+2)} \sum_{s=0}^{n} I_{n,r}(s),$$

where

$$(16)\qquad I_{n,r}(s) = \int u^r h_s(u)h_{n-s}(u)\,du \qquad (0 \leqq s \leqq n).$$

For $I_{n,0}(s)$ we obtain on applying reduction formula (9)

$$I_{n,0}(s) = \int h_{n-s}(u)h_s(u)\,du = \int h_{n-s}(u)\int \phi(u-y)h_{s-1}(y)\,dy\,du \qquad (s \geqq 1)$$

and, reversing the order of integration and using (9) again,

$$I_{n,0}(s) = \int h_{s-1}(y)\int \phi(u-y)h_{n-s}(u)\,du\,dy$$

$$= \int h_{s-1}(y)h_{n-s+1}(y)\,dy = I_{n,0}(s-1) \qquad (1 \leqq s \leqq n).$$

Hence, using (12) and (7),

$$I_{n,0}(s) = I_{n,0}(0) = \int h_0(y)h_n(y)\,dy \tag{17}$$

$$= h_{n+1} = (2\pi)^{-1/2}(n+2)^{-3/2} \qquad (0 \leqq s \leqq n, n \geqq 0).$$

Substituting this result into (15) and noting (4), we obtain

$$\mu_0'(n+2) = \frac{n+1}{n+2} = 1 - P_{n+2}(0)$$

for the zero order moment, as is to be expected when one recalls that $P_n(u)$ is zero for $u < 0$ and has a saltus at $u = 0$.

For $I_{n,1}(s)$ we proceed in the same manner but after reversing the order of integration we apply the first of the two reduction formulae (10) for the derivative to obtain

$$I_{n,1}(s) - I_{n,1}(s-1) = J_n(s-1) \qquad (1 \leqq s \leqq n), \tag{18}$$

where

$$J_n(s) = \int h_s(u)h'_{n-s}(u)\,du \qquad (0 \leqq s \leqq n).$$

Similarly we obtain a difference equation for $J_n(s)$ by applying the reduction formula (9) to $h_s(u)$ in the integrand of $J_n(s)$, reversing the order of integration, applying the second form of formula (10) to $h'_{n-s}(u)$, and simplifying the result by using (12). The resulting difference equation is

$$J_n(s) - J_n(s-1) = -h_s h_{n-s} \qquad (1 \leqq s \leqq n)$$

which when summed over s gives

$$J_n(t) = J_n(0) - \sum_{s=1}^{t} h_s h_{n-s} \qquad (1 \leqq t \leqq n).$$

By Lemma 4 $J_n(0) = I_{n,1}(0) - h_0 h_n$ so we may write

$$J_n(s) = I_{n,1}(0) - \sum_{s=0}^{s} h_s h_{n-s} \qquad (0 \leqq s \leqq n). \tag{19}$$

Returning to (18) we substitute (19) for $J_n(s)$ and sum this difference equation for $I_{n,1}(s)$ over s from 1 to t, reversing the order of summation to obtain

$$I_{n,1}(t) = (t+1)I_{n,1}(0) - \sum_{v=0}^{t-1} (t-v)h_v h_{n-v} \qquad (1 \leqq t \leqq n). \tag{20}$$

We note that this expression is also valid for $t = 0$ if we write the range of summation for v from 0 to t. Since $I_{n,1}(s) = I_{n,1}(n-s)$ by definition (16), we have upon putting $t = n$ in this extended form of (20)

$$I_{n,1}(0) = (n+1)I_{n,1}(0) - \sum_{v=0}^{n} (n-v)h_v h_{n-v}.$$

But

$$\sum_{v=0}^{n} vh_v h_{n-v} = \sum_{v=0}^{n} (n-v)h_v h_{n-v} = n\sum_{v=0}^{n} h_v h_{n-v} - \sum_{v=0}^{n} vh_v h_{n-v},$$

hence

$$\sum_{v=0}^{n} (n-v)h_v h_{n-v} = \frac{n}{2}\sum_{v=0}^{n} h_v h_{n-v}$$

and

$$I_{n,1}(0) = \frac{1}{2}\sum_{v=0}^{n} h_v h_{n-v}. \tag{21}$$

Substituting (21) into the extended form of (20) and summing over t, we have after reversing the order of summation

$$\sum_{t=0}^{n} I_{n,1}(t) = \frac{(n+1)(n+2)}{4}\sum_{v=0}^{n} h_v h_{n-v} - \sum_{v=0}^{n} \frac{(n-v)(n-v+1)}{2} h_v h_{n-v} \qquad (n \geqq 0).$$

In this last expression we note that the coefficient of $h_v h_{n-v}$ is

$$2\frac{(n+1)(n+2)}{4} - \frac{(n-v)(n-v+1)}{2} - \frac{v(v+1)}{2} = (v+1)(n-v+1) \qquad (0 \leqq v \leqq n)$$

so that, using Lemma 1, we may write

$$\sum_{s=0}^{n} I_{n,1}(s) = \frac{1}{2}\sum_{s=0}^{n} (s+1)(n-s+1)h_s h_{n-s} = \frac{1}{4\pi}\sum_{s=1}^{n+1} s^{-1/2}(n-s+2)^{-1/2} \qquad (n \geqq 0).$$

Substituting this into (15) we obtain

$$\mu_1'(n+2) = \frac{1}{2}\sqrt{\frac{n+2}{2\pi}} \sum_{s=1}^{n+1} s^{-1/2}(n+2-s)^{-1/2} \qquad (n \geqq 0) \tag{22}$$

for the first moment of the maximum and, hence, twice this value for the mean range of the adjusted partial sums.

For $I_{n,2}(s)$ the method is similar though more tedious. Applying (9) to $h_s(u)$ in the integrand of $I_{n,2}(s)$, reversing the order of integration, and applying the first form of (11) to the inner integral we have

$$I_{n,2}(s) = I_{n,2}(s-1) + K_n(s-1) + 2L_n(s-1) + I_{n,0}(s-1) \quad (1 \leqq s \leqq n), \tag{23}$$

where

$$\left.\begin{aligned} K_n(s) &= \int h_s(u) h_{n-s}''(u)\, du, \\ L_n(s) &= \int u h_s(u) h_{n-s}'(u)\, du \end{aligned}\right\} \qquad (0 \leqq s \leqq n).$$

We obtain, in the same way as was done for $J_n(s)$ [but using the second and third forms of (11)] difference equations for $L_n(s)$ and $K_n(s)$ respectively

$$\left.\begin{aligned} L_n(s) - L_n(s-1) &= K_n(s-1), \\ K_n(s) - K_n(s-1) &= h_{n-s}h_s' - h_s h_{n-s}' \end{aligned}\right\} \qquad (1 \leqq s \leqq n).$$

Using Lemma 4 to evaluate $L_n(0)$ and $K_n(0)$, we find the solutions of these equations to be, for $0 \leqq s \leqq n$,

$$L_n(s) = (s+1)[I_{n,2}(0) - I_{n,0}(0)] + \sum_{v=0}^{s} (s-v)[h_{n-v}h_v' - h_v h_{n-v}'],$$

$$K_n(s) = I_{n,2}(0) - I_{n,0}(0) + \sum_{v=0}^{s} [h_{n-v}h_v' - h_v h_{n-v}'].$$

Substituting these expressions into (23) we have after summing over s and rearranging

$$I_{n,2}(t) = (t+1)^2 I_{n,2}(0) - t(t+1) I_{n,0}(0) + \sum_{v=0}^{t} (t-v)^2 [h_{n-v}h_v' - h_v h_{n-v}'] \qquad (0 \leqq t \leqq n). \tag{24}$$

Since $I_{n,2}(s) = I_{n,2}(n-s)$, we now evaluate $I_{n,2}(0)$ as before obtaining

$$I_{n,2}(0) = \frac{n+1}{n+2} I_{n,0}(0) + \frac{1}{n+2} \sum_{v=0}^{n} v[h_{n-v}h_v' - h_v h_{n-v}'],$$

after we have observed that, as for $I_{n,1}(0)$, the summations involved satisfy

certain identities, in particular

$$\sum_{v=0}^{n}(n-v)^2[h_{n-v}h'_v - h_v h'_{n-v}] = -n\sum_{v=0}^{n} v[h_{n-v}h'_v - h_v h'_{n-v}].$$

Substituting the expression for $I_{n,2}(0)$ into (24) and summing over t, we have

$$\sum_{t=0}^{n} I_{n,2}(t) = \frac{(n+1)(n+3)}{6} I_{n,0}(0) + \frac{1}{6}\sum_{v=0}^{n}[(n+1)(2n+3)v + (n-v)(n-v+1)(2n-2v+1)][h_{n-v}h'_v - h_v h'_{n-v}]$$

$$= \frac{(n+1)(n+3)}{6} I_{n,0}(0) + \frac{1}{3}\sum_{v=1}^{n}(v+1)(n-v+1)(2v-n)h_{n-v}h'_v.$$

By (13) $h'_v = I_{v-1,1}(0)$ so that, using (7), (17) and (21), we may write

$$\sum_{t=0}^{n} I_{n,2}(s) = \frac{1}{6}\left[\frac{(n+1)(n+3)}{\sqrt{2\pi(n+2)^3}} + \frac{1}{(2\pi)^{3/2}}\sum_{s=2}^{n+1}\sum_{t=1}^{s-1}\frac{s(2s-n-2)}{\sqrt{(n+2-s)t^3(s-t)^3}}\right] \quad (n \geqq 0)$$

provided we interpret the summation as zero when $n = 0$. Hence from (15) the second moment of the maximum is

(25)
$$\mu'_2(n+2) = \frac{1}{6}\cdot\left[\frac{(n+1)(n+3)}{n+2} + \frac{\sqrt{n+2}}{2\pi}\sum_{s=2}^{n+1}\sum_{t=1}^{s-1}\frac{s(2s-n-2)}{\sqrt{(n+2-s)t^3(s-t)^3}}\right] \quad (n \geqq 0).$$

A table of values of μ'_1, μ'_2 and σ for samples of size 10, 20, $\cdots$, 150, was computed from formulae (22) and (25), (see Table 1).

5. Asymptotic values for the first and second moments. Feller [2] has considered the asymptotic distribution of the adjusted range for large n and found the asymptotic value of the mean adjusted range to be $[(\pi/2)n]^{1/2}$. That this is in agreement with our result can be seen by approximating to the sum in our formula (22)

$$\mu'_1(n) = \frac{1}{2}\sqrt{\frac{n}{2\pi}}\sum_{s=1}^{n-1} s^{-1/2}(n-s)^{-1/2}$$

by the integral

$$\int_1^{n-1} s^{-1/2}(n-s)^{-1/2}\,ds.$$

On making the substitution $n\theta = s$, this integral becomes

$$\int_{1/n}^{1-1/n} \theta^{-1/2}(1-\theta)^{-1/2}\,d\theta,$$

which approaches $B(\frac{1}{2}, \frac{1}{2}) = \pi$ as n becomes large. Thus the asymptotic value for the mean of the maximum of the adjusted partial sums is

$$\mu_1'(n) \sim \frac{1}{2}\sqrt{\frac{\pi n}{2}} \doteq 0.6267 n^{1/2} \tag{26}$$

and the asymptotic mean range is $[(\pi/2)n]^{1/2}$ as obtained by Hurst and Feller.

Similarly we obtain the asymptotic value of the second moment of the maximum from our formula (25)

$$\mu_2'(n) = \frac{1}{6}\left[\frac{n^2-1}{n} + \frac{\sqrt{n}}{2\pi}\sum_{s=2}^{n-1}\sum_{t=1}^{s-1}\frac{s(2s-n)}{\sqrt{(n-s)t^3(s-t)^3}}\right]$$

by approximating to the double sum by the double integral

$$I = \int_{s=2}^{n-1}\int_{t=1}^{s-1}\frac{s(2s-n)}{\sqrt{(n-s)t^3(s-t)^3}}\,dt\,ds.$$

Integrating first with respect to t by means of the substitution $t = (sz)^{-1}$, we have

$$I = \tfrac{1}{4}\int_2^{n-1}\frac{(2s-n)(s-2)}{s\sqrt{(n-s)(s-1)}}\,ds =$$

$$\tfrac{1}{4}\int_2^{n-1}\left[\sqrt{\frac{s-1}{n-s}} - \sqrt{\frac{n-s}{s-1}} - \frac{3}{\sqrt{(n-s)(s-1)}} + \frac{2n}{s\sqrt{(n-s)(s-1)}}\right]ds.$$

Applying the substitution $s = n\sin^2\theta + \cos^2\theta$ to the first three terms and the substitution $s = 2n[n+1+(n-1)\sin\theta]^{-1}$ to the last term of this integrand,

TABLE 1

Values of $\mu_1'(n)$, $\mu_2'(n)$, σ_n and the asymptotic approximations for $\mu_1'(n)$ and σ_n

n	Exact values			Asymptotic approximation	
	$\mu_1'(n)$	$\mu_2'(n)$	σ_n	$\mu_1' \sim 0.6267\ n^{1/2}$	$\sigma_n \sim 0.3276\ n^{1/2}$
10	1.3948	3.019	1.0358	1.9817	1.0359
20	2.2178	7.068	1.4660	2.8025	1.4649
30	2.8483	11.336	1.7952	3.4323	1.7942
40	3.3796	15.718	2.0726	3.9633	2.0717
50	3.8477	20.173	2.3170	4.4311	2.3162
60	4.2707	24.680	2.5378	4.8541	2.5373
70	4.6597	29.226	2.7409	5.2430	2.7406
80	5.0218	33.803	2.9298	5.6050	2.9298
90	5.3619	38.405	3.1072	5.9450	3.1076
100	5.6835	43.028	3.2749	6.2666	3.2757
110	5.9894	47.667	3.4342	6.5724	3.4356
120	6.2817	52.322	3.5863	6.8647	3.5883
130	6.5620	56.989	3.7321	7.1450	3.7349
140	6.8318	61.667	3.8721	7.4147	3.8758
150	7.0920	66.351	4.0067	7.6749	4.0118

we obtain

$$I = -24 \sin^{-1}\left(1 - \frac{2}{n-1}\right) + 8\sqrt{n} \sin^{-1}\left[1 - \frac{2}{(n-1)^2}\right] - 8\sqrt{n} \sin^{-1}\left(\frac{1}{n-1}\right),$$

which approaches $-12\pi + 4\pi\sqrt{n}$ for large n. Thus the asymptotic value for the second moment about the origin is

$$\mu_2'(n) \sim \frac{n}{2} - \sqrt{n}$$

and, using the asymptotic value obtained for $\mu_1'(n)$ in (26), we find the asymptotic value of the variance of the maximum of the adjusted partial sums is

$$\sigma_n^2 \sim \left(\frac{1}{2} - \frac{\pi}{8}\right) n \doteq 0.1073n. \tag{27}$$

Comparing this asymptotic value with that obtained by Anis [4] for the variance of the maximum of the unadjusted sums which was $[1 - (2/\pi)]n \doteq 0.3634n$, we see that Feller's comment on the greater stability of the adjusted partial sums is well borne out by our results.

In Table 1 we note that:

1) the series for $\mu_1'(n)$ converges very slowly so that the asymptotic approximation (26) should not be used for values of n within the range of this table,
2) the series for $\mu_2'(n)$ converges even more slowly, but
3) the asymptotic approximation (27) gives very good values for σ_n even within the range of the table because the errors in the approximations to $\mu_1'(n)$ and $\mu_2'(n)$ are in the same direction and largely cancel.

REFERENCES

[1] H. E. Hurst, "Long-term storage capacity of reservoirs," *Trans. Amer. Soc. Civil Engrs.*, Vol. 116 (1951), p. 770.

[2] W. Feller, "The asymptotic distribution of the range of sums of independent random variables," *Ann. Math. Stat.*, Vol. 22 (1951), pp. 427–432.

[3] A. A. Anis and E. H. Lloyd, "On the range of partial sums of a finite number of independent normal variates," *Biometrika*, Vol. 40 (1953), pp. 35–42.

[4] A. A. Anis, "The variance of the maximum of partial sums of a finite number of independent normal variates," *Biometrika*, Vol. 42 (1955), pp. 96–101.

[5] A. A. Anis, "On the moments of the maximum of partial sums of a finite number of independent normal variates," *Biometrika*, Vol. 43 (1956), pp. 79–84.

REMARKS CONCERNING CHARACTERISTIC FUNCTIONS

By Eugene Lukacs

Office of Naval Research[1]

1. Summary. In the first part of this note, we study functions of characteristic functions which are themselves characteristic functions and discuss also a property of analytic characteristic functions. In the second part, an example is constructed to answer a question raised by D. Dugué [3].

2. Functions of characteristic functions. Let $F(x)$ be a distribution function, that is, a never-decreasing function which is continuous to the right and is such that $F(-\infty) = 0$ while $F(+\infty) = 1$. Its Fourier transform

$$\phi(t) = \int_{-\infty}^{\infty} e^{itx}\, dF(x) \tag{1}$$

is called the characteristic function of the distribution $F(x)$. Characteristic functions are very important in probability theory, and in the following discussion we shall use some of their well-known properties which may be found in books [1], [5] on the subject. We first derive a theorem which shows how given characteristic functions may be transformed into new characteristic functions.

Theorem 1. *Let $\{\phi_\nu(t)\}$ be an arbitrary sequence of characteristic functions and $\{a_\nu\}$ be a sequence of real numbers. The necessary and sufficient condition that*

$$f(t) = \sum_{\nu=0}^{\infty} a_\nu \phi_\nu(t) \tag{2}$$

should be a characteristic function for every sequence $\{\phi_\nu(t)\}$ of characteristic functions is that

$$a_\nu \geqq 0, \qquad \sum_{\nu=0}^{\infty} a_\nu = 1. \tag{3}$$

We first show that the condition is sufficient. Let m ($m \geqq 0$) be a subscript such that a_m is the first non-vanishing element of the sequence $\{a_\nu\}$. We denote by

$$g_n(t) = \left[\sum_{\nu=0}^{m+n} a_\nu \phi_\nu(t)\right] \Big/ \left[\sum_{\nu=0}^{m+n} a_\nu\right] \qquad \text{for } n = 0, 1, 2, \cdots.$$

If (3) is satisfied, then $g_n(t)$ is a linear combination of a finite number of characteristic functions. The coefficients in this linear combination are non-negative and their sum is one; therefore $g_n(t)$ is also a characteristic function. We see

Received June 17, 1955.

[1] Now at The Catholic University of America.

then from P. Lévy's continuity theorem that $f(t) = \lim_{n\to\infty} g_n(t)$ is also a characteristic function.

To prove the necessity of the condition, we assume that $f(t)$ as given by (2) is a characteristic function for any sequence $\phi_\nu(t)$ of characteristic functions. Let $\phi_\nu(t) = e^{it\nu}$; then $f(t) = \sum_{\nu=0}^{\infty} a_\nu e^{it\nu}$. This is the Fourier transform of a step function with jumps a_ν at the points $\nu = 0, 1, 2, \cdots$. Since $f(t)$ is by assumption a characteristic function, this step function must be a discrete probability distribution; therefore $a_\nu \geqq 0$, $\sum a_\nu = 1$, so that Theorem 1 is established.

An application of some interest is obtained by putting $\phi_n(t) = n^{-it} = \exp[-it(\ln n)]$ and $a_n = n^{-\sigma} / \sum_1^{\infty} n^{-\sigma}$, where $\sigma > 1$. It follows then from Theorem 1 that the function $f(t) = \zeta(\sigma + it)/\zeta(\sigma)$ is a characteristic function for $\sigma > 1$. Here $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$ is Riemann's zeta function and $s = \sigma + it$, with σ, t real and $\sigma > 1$.

This result was already obtained in a different manner by B. V. Gnedenko and A. N. Kolmogorov ([7], p. 75) who showed that $\zeta(\sigma + it)/\zeta(\sigma)$ is the characteristic function of an infinitely divisible distribution.

Next, we let $\phi(t)$ be an arbitrary characteristic function and put $\phi_\nu(t) = [\phi(t)]^\nu$, $\nu = 0, 1, 2, \cdots$. We obtain then

COROLLARY TO THEOREM 1. *Let $\phi(t)$ be a characteristic function and let $G(z)$ be a function of the complex variable z, which is regular in $|z| < R$ where $R > 1$. The function $G[\phi(t)]$ is also a characteristic function if, and only if, $G(z)$ has a power-series expansion about the origin with non-negative coefficients and if $G(1) = 1$.*

It is worth while to remark that the class of functions $G(z)$ which have the property that $G[\phi(t)]$ is a characteristic function whenever $\phi(t)$ is a characteristic function includes also functions which are not analytic. An example is the function $G(z) = |z|^2$. The restriction in the corollary that $G(z)$ should be regular is therefore somewhat artificial.

DEFINITION. A distribution is said to be infinitely divisible if for every positive integer n its characteristic function is the nth power of some characteristic function.

By means of the corollary to Theorem 1 we obtain the following result.

THEOREM 2. *Let $\phi(t)$ be an arbitrary characteristic function and p a real number such that $p > 1$; then*

$$\psi(t) = \frac{p - 1}{p - \phi(t)} \tag{4}$$

is the characteristic function of an infinitely divisible law.

To prove Theorem 2 we let n be a positive integer and consider the function

$$G(z) = \left[\frac{p - 1}{p - z}\right]^{1/n}.$$

Here it is understood that $G(z)$ is the principal value of the power on the right-hand side. Clearly

$$G(z) = \left[\frac{p-1}{p}\right]^{1/n}\left[\frac{1}{1-z/p}\right]^{1/n} = \left[\frac{p-1}{p}\right]^{1/n}\left[1-\frac{z}{p}\right]^{-1/n}.$$

We expand $G(z)$ according to the binomial theorem and see that

$$G(z) = \left[\frac{p-1}{p}\right]^{1/n}\left\{1 + \sum_{k=1}^{\infty} \frac{(1+n)(1+2n)\cdots(1+\overline{k-1}n)}{(np)^k k!} z^k\right\}.$$

This shows that for any positive integer n the conditions of the corollary are satisfied. The function $G[\phi(t)] = \{[p-1]/[p-\phi(t)]\}^{1/n}$ is therefore a characteristic function for any positive integer n; in other words, $\psi(t)$, as given by (4), is the characteristic function of an infinitely divisible law.

In a similar manner we derive from the corollary to Theorem 1 a theorem which is due to Bruno de Finetti [4].

THEOREM OF DE FINETTI. *If $\phi(t)$ is an arbitrary characteristic function, and if p is a positive real number, then $\psi(t) = \exp\{p[\phi(t) - 1]\}$ is the characteristic function of an infinitely divisible law.*

The function $G(z) = e^{p(z-1)}$ satisfies the assumptions of the corollary, so that we see immediately that $\psi(t)$ is a characteristic function for any $p > 0$. It follows then from its functional form that it must be the characteristic function of an infinitely divisible law.

3. A remark concerning analytic characteristic functions. A characteristic function is said to be an analytic characteristic function if it is an analytic function which coincides in some neighborhood of the origin with a characteristic function.

In an earlier paper [6] the following result was obtained:

THEOREM 4 of [6]. *Let $\phi(t)$ be the characteristic function of an infinitely divisible law and assume that $\phi(t)$ is an analytic characteristic function. Then $\phi(t)$ has no zeros inside its strip of convergence.*

In the following we show that this statement cannot be improved. This is done by constructing an analytic characteristic function of an infinitely divisible law which has zeros on the boundary of its strip of convergence.

Let $a > 0$, $b > 0$ be two real numbers and put $w = a + ib$; the function

$$\phi(t) = \frac{(1 - it/w)(1 - it/\overline{w})}{(1 - it/a)^2} \tag{5}$$

is then a characteristic function. This is seen immediately if we write $\phi(t) = p + (1-p)(1-it/a)^{-2}$, where $p = a^2/(a^2+b^2)$. We define

$$m = 2\int_0^{\infty} \frac{e^{-ax}(1-\cos bx)}{1+x^2}\,dx \tag{6a}$$

$$M(x) = \begin{cases} 0 & \text{for } x < 0 \\ -2\displaystyle\int_x^{\infty} e^{-at}(1-\cos bt)t^{-1}\,dt & \text{for } x > 0. \end{cases} \tag{6b}$$

The function $M(x)$ is real and non-decreasing in $(-\infty, 0)$ and in $(0, \infty)$. More-

over, $M(-\infty) = M(+\infty) = 0$; and the integral $\int u^2\, dM(u)$ is finite over every finite interval. According to P. Lévy's representation theorem ([5] p. 180), the function

$$\psi(t) = mit + \int_{-\infty}^{\infty} \left(e^{itx} - 1 - \frac{itx}{1+x^2} \right) dM(x) \tag{7}$$

is the logarithm of the characteristic function of an infinitely divisible law. We write

$$I(t) = \int_0^{\infty} \left(e^{itx} - 1 - \frac{itx}{1+x^2} \right) e^{-\alpha x}(1 - \cos bx)\, \frac{dx}{x}$$

and obtain $\psi(t) = mix + 2I(t)$.

It is easily seen that it is permissible to differentiate $I(t)$ under the integral sign. A simple computation gives

$$2I'(t) = 2\, \frac{i/a}{1 - it/a} - \frac{i/w}{1 - it/w} - \frac{i/\overline{w}}{1 - it/\overline{w}} - im.$$

Considering $I(0) = 0$, we see that $\psi(t) = \log \phi(t)$, where $\phi(t)$ is given by (5).

We finally remark that it is possible to use Theorem 2 to construct characteristic functions of infinitely divisible laws which have zeros arbitrarily close to the boundary of the strip of convergence. As an example we mention $\psi(t) = (p - 1)/[p - \phi(t)]$, where $\phi(t) = (1 - it/\alpha)^{-1}$. The function $\psi(t)$ then has the zero $t_0 = -i\alpha$, and the boundary of its region of convergence is the line $\operatorname{Im}(t) = -\alpha(p - 1)/p$. By selecting p large enough, the distance between the boundary and t_0 can be made arbitrarily small.

4. A question raised by D. Dugué. In this section we are concerned with certain factorizations of non-infinitely divisible laws. The uniform (rectangular) distribution has the characteristic function $(\sin t)/t$; it is not infinitely divisible, since it has real zeros. It is well known that it has the factor $\sin (t/n)/(t/n)$ for every positive integer n. The uniform distribution is therefore an example of a law which is not infinitely divisible but has an enumerable infinity of different factors. These factors, with characteristic function $\sin (t/n)/(t/n)$, depend on a discrete parameter n.

In a recent paper [3], D. Dugué raised the question of whether there exists a law which is not infinitely divisible but has a non-enumerable set of factors depending on a continuous parameter. As an example for such a distribution Dugué uses the Laplace distribution. This example is, however, invalid, since the Laplace distribution is infinitely divisible. The purpose of this section is to answer Dugué's question in the affirmative by giving an example of a probability law with the desired properties. This example will be a rational characteristic function; for its construction we use the following lemma.

LEMMA 1. *Let*

$$\phi(t) = \frac{\left(1+\frac{it}{w}\right)\left(1+\frac{it}{\bar{w}}\right)}{\left(1-\frac{it}{a}\right)\left(1-\frac{it}{v}\right)\left(1-\frac{it}{\bar{v}}\right)}, \tag{9}$$

where $v = a + ib$, $w = \alpha + i\beta$, *and* $a > 0$, $b > 0$, $\alpha > 0$, $\beta > 0$. *The function* $\phi(t)$ *is a characteristic function if, and only if, one of the following two, mutually exclusive, conditions holds*:

(i) $\beta = \sqrt{b^2 - (a+\alpha)^2} \geqq \sqrt{3}(a+\alpha)$;
(ii) $\beta \neq \sqrt{b^2 - (a+\alpha)^2}$ *and simultaneously* $\beta^2 \geqq (a+\alpha)^2 + b^2/2$.

PROOF. We denote by

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi(t)\,dt$$

and obtain by a simple computation

$$f(x) = \begin{cases} Ce^{-\alpha x}\left\{1 - \left[\frac{c^2 - b^2}{c^2}\right]\cos bx - \frac{2bd}{c^2}\sin bx\right\} & \text{if } x > 0; \\ 0 \quad \text{otherwise.} \end{cases} \tag{10}$$

Here

$$\begin{cases} d = a + \alpha, \\ c^2 = (a+\alpha)^2 + \beta^2 = d^2 + \beta^2, \\ C = \dfrac{av\bar{v}c^2}{w\bar{w}b^2}. \end{cases} \tag{10a}$$

The function $f(x)$ is real and $\int_{-\infty}^{\infty} f(x)\,dx = 1$. Therefore we conclude that the function $\phi(t)$ is a characteristic function if, and only if, the trigonometric polynomial

$$h(x) = 1 - \left[\frac{c^2 - b^2}{c^2}\right]\cos bx - \frac{2bd}{c^2}\sin bx \tag{11}$$

is nonnegative.

We assume first that $c^2 = b^2$ or, considering (10a), that $\beta^2 = b^2 - (a+\alpha)^2$. Then $h(x) = 1 - 2d/b \sin bx$ and $h(x) \geqq 0$ if, and only if, $2d/b \leqq 1$. It is seen by simple algebra that this is equivalent to condition (i). We suppose next that $c^2 \neq b^2$ and determine by elementary considerations the smallest value of $h(x)$, which is

$$\min_{-\infty < x < +\infty} h(x) = 1 - \frac{|c^2 - b^2|}{c^2} \cdot \frac{1}{\cos bx_0},$$

where $-\pi/2b < x_0 < +\pi/2b$ and $\tan bx_0 = 2bd/(c^2 - b^2)$. The function $h(x)$ is therefore non-negative if, and only if, $1 \geqq |c^2 - b^2|/(c^2 \cos bx_0)$. This condition leads easily to (ii), so that the lemma is established.

We use this lemma, together with the theorem quoted at the beginning of Section 3, to construct a characteristic function with the desired properties.

Let a, α_1, α_2, β_1, β_2, b be arbitrary positive numbers such that $\alpha_2 > \alpha_1$ and also

$$\beta_j^2 > \max\left[b^2 - (a + \alpha_j)^2;\quad (a + \alpha_j)^2 + \frac{b^2}{2}\right] \qquad (j = 1, 2). \tag{12}$$

We define $v = a + ib$, $w_1 = \alpha_1 + i\beta_1$, and $w_2 = \alpha_2 + i\beta_2$. The functions

$$\phi_j(t) = \frac{\left(1 + \frac{it}{w_j}\right)\left(1 + \frac{it}{\overline{w}_j}\right)}{\left(1 - \frac{it}{a}\right)\left(1 - \frac{it}{v}\right)\left(1 - \frac{it}{\bar{v}}\right)} \qquad (j = 1, 2)$$

are then characteristic functions, since they satisfy condition (ii) of the preceding lemma. Therefore

$$\phi_3(t) = \phi_1(t)\phi_2(t) = \frac{\left(1 + \frac{it}{w_1}\right)\left(1 + \frac{it}{\overline{w}_1}\right)\left(1 + \frac{it}{w_2}\right)\left(1 + \frac{it}{\overline{w}_2}\right)}{\left(1 - \frac{it^2}{a}\right)\left(1 - \frac{it}{v}\right)^2\left(1 - \frac{it}{v}\right)^2} \tag{13}$$

is also a characteristic function. It is also known that

$$\phi_4(t) = \left[\left(1 + \frac{it}{\alpha_2}\right)\left(1 + \frac{it}{w_2}\right)\left(1 + \frac{it}{\overline{w}_2}\right)\right]^{-1}$$

is a characteristic function; we conclude then that this is also true for $\phi(t) = \phi_4(t)\phi_3(t)$; i.e.,

$$\phi(t) = \frac{\left(1 + \frac{it}{w_1}\right)\left(1 + \frac{it}{\overline{w}_1}\right)}{\left(1 + \frac{it}{\alpha_2}\right)\left(1 - \frac{it}{a}\right)^2\left(1 - \frac{it}{v}\right)^2\left(1 - \frac{it}{\bar{v}}\right)^2}. \tag{14}$$

The function $\phi(t)$ is an analytic characteristic function which has the strip $\alpha_2 > I(t) > -a$ as its strip of convergence. Its zeros iw_1 and $i\overline{w}_1$ are located inside this strip, so that $\phi(t)$ cannot be the function of an infinitely divisible law. Similarly, $\phi_3(t)$ is not infinitely divisible, since its strip of convergence is the half plane $I(t) > -a$ in which it has four zeros. We have, therefore, an example of a law $\phi(t)$, which is not infinitely divisible and which has, nevertheless, a non-enumerable infinity of not infinitely divisible factors $\phi_3(t)$. These factors depend on a continuous parameter β_2, which is subject only to the restriction (12).

REFERENCES

[1] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, N. J., 1946.

[2] H. Cramér, "On the factorization of certain probability distributions," *Ark. Mat.*, Vol. 1 (1949), pp. 61–65.

[3] D. Dugué, "Sur certains exemples de décompositions en arithmétique des lois de probabilité," *Ann. Inst. H. Poincaré*, Vol. 12 (1951), pp. 159–169.

[4] B. de Finetti, "Le funzioni caratteristiche di legge instantanea," *Rend. Accad. Lincei*, Ser. 6, Vol. 12 (1930), pp. 278–282.

[5] P. Lévy, *Théorie de l'Addition des Variables Aléatoires*, Gauthier-Villars & Cie, Paris, 1937.

[6] E. Lukacs and O. Szász, "Analytic characteristic functions," *Pacific J. Math.*, Vol. 2 (1952), pp. 615–625.

[7] B. V. Gnedenko and A. N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables* (trans. by K. L. Chung), Addison Wesley Publishing Co., Cambridge, Mass., 1954.

THE DISTRIBUTION OF THE RATIOS OF CERTAIN QUADRATIC FORMS IN TIME SERIES[1]

By Seymour Geisser[2]

University of North Carolina

1. Introduction. In testing the hypothesis that successive members of a series of observations are serially correlated a number of statistics have been proposed. Durbin and Watson [4] gave the exact distribution of several of these statistics when they are slightly modified. We shall extend the work of Durbin and Watson for a non-null case of two of their modified statistics and also find a simple expression for the moments of another of their statistics.

2. The Double root result. Assume that $X' = (x_1, x_2, \cdots, x_n)$ has probability density

$$f(X) = |\Lambda|^{1/2}(2\pi)^{-n/2} \exp[-X'\Lambda X/2], \tag{2.1}$$

where Λ is a positive definite matrix and $n = 2m$. Let

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_1 \end{pmatrix}; \quad B = \begin{pmatrix} B_1 & 0 \\ 0 & B_1 \end{pmatrix}; \quad \Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_1 \end{pmatrix}; \tag{2.2}$$

where B_1 is positive definite or positive semi-definite and of rank $m - q$ which is $\geqq$ the rank of A_1, a real symmetric matrix. Further assume that A, B, and Λ commute pairwise, and that the characteristic roots a_j of A and the characteristic roots b_j of B are so numbered that if $a_j \gtrless 0, b_j > 0$ and $a_j/b_j \geqq a_{j+1}/b_{j+1}$ for all a_j and a_{j+1} which are $\neq 0$.

Now

$$G(z) = P\left[\frac{X'AX}{X'BX} \leqq z\right] = P[X'(A - zB)X \leqq 0], \tag{2.3}$$

where X is $N(0, \Lambda^{-1})$. Making an orthogonal transformation $X = PY$ where $P'AP = D_a$, $P'BP = D_b$, $P'\Lambda P = D_\lambda$ are diagonal matrices with elements $a_j = a_{m+j}$, $b_j = b_{m+j}$ and $\lambda_j = \lambda_{m+j}$, we get

$$G(z) = P[Y'(D_a - zD_b)Y \leqq 0], \tag{2.4}$$

where Y is $N(0, D_\lambda^{-1})$. Now let $Y = D_\lambda^{-1/2}W$ so that

$$G(z) = P[W'(D_a - zD_b)D_\lambda^{-1}W \leqq 0], \tag{2.5}$$

Received June 28, 1955; revised January 3, 1957.

[1] Work under contract with the Office of Naval Research NR 042 031, for investigations in statistics and probability at Chapel Hill. Reproduction in whole or in part is permitted for any purpose of the United States Government.

[2] Now with the National Institute of Mental Health.

where W is $N(0, I)$. Hence by the duplication of roots a_j, b_j and λ_j we get

$$G(z) = P\left[\sum_{j=1}^{m-q} \lambda_j^{-1}(a_j - zb_j)w_j^2 \leqq 0\right] \tag{2.6}$$

where w_j^2 are independent and each is a χ^2 variable with 2 degrees of freedom.

Using a result by R. L. Anderson [1] which in general terms states that if the c_j are all different then

$$P\left[\sum_{j=1}^{p} c_j w_j^2 \leqq 0\right] = 1 - \sum_{j \in s} c_j^{p-1} \prod_{\substack{j=1 \\ j \neq k}}^{p} (c_k - c_j)^{-1} \tag{2.7}$$

where $S = \{j/c_j > 0\}$, we find that

$$\begin{aligned} G(z) = 1 - \prod_{j=1}^{m-q} \lambda_j \sum_{k=1}^{L} (a_k - b_k z)^{m-q-1} \lambda_k^{-1} \\ \cdot \prod_{\substack{j=1 \\ j \neq k}}^{m-q} [\lambda_j(a_k - b_k z) - \lambda_k(a_j - b_j z)]^{-1}, \end{aligned} \tag{2.8}$$

where

$$\frac{a_{L+1}}{b_{L+1}} \leqq z \leqq \frac{a_L}{b_L}$$

for $L = 1, \cdots, m - q$. This result could also have been gotten by contour integration through the results of Gurland [6] or by Madow's generalization of Anderson's result.

3. The distribution of the Durbin and Watson Statistic in the non-null case. Using the result (2.8) and letting

$$R = \sum_{\substack{i=1 \\ i \neq m}}^{2m-1} x_{i+1} x_i \Big/ \sum_{i=1}^{2m} x_i^2, \tag{3.1}$$

where

$$\Lambda_1 = \begin{pmatrix} 1 + \rho^2 & & & -\rho \\ -\rho & \cdot & & \\ & & \cdot & \\ & & \cdot & -\rho \\ & & -\rho & 1 + \rho^2 \end{pmatrix}, \tag{3.2}$$

we may find the distribution of R. For $a_j = \cos(j\pi)/(m + 1)$, $\lambda_j = 1 + \rho^2 - 2\rho a_j$, $b_j = 1$. Now $\prod_{j=1}^{m} \lambda_j = (1 - \rho^{2m+2})(1 - \rho^2)^{-1}$ and $\lambda_k - \lambda_j = (a_k - a_j) \cdot (1 + \rho^2 - 2\rho z)$ and by Geisser [5]

$$\prod_{\substack{j=1 \\ j \neq k}}^{m} (a_k - a_j) = (m + 1)(-1)^{k+1} 2^{-m} \csc^2 \frac{k\pi}{m + 1}. \tag{3.3}$$

Therefore

$$G(z;\rho) = 1 - (1-\rho^{2m+2})2^m(m+1)^{-1}(1-\rho^2)^{-1}(1+\rho^2-2\rho z)^{1-m}$$

$$\cdot \sum_{k=1}^{L}(-1)^{k+1}\left(\cos\frac{k\pi}{m+1}-z\right)^{m-1}\sin^2\frac{k\pi}{m+1}\left(1+\rho^2-2\rho\cos\frac{k\pi}{m+1}\right)^{-1} \tag{3.4}$$

and

$$G'(z;\rho) = g(R;\rho)$$

$$= (1-\rho^{2m+2})2^m(m-1)(m+1)^{-1}(1-\rho^2)^{-1}(1+\rho^2-2\rho R)^{-m} \tag{3.5}$$

$$\cdot \sum_{k=1}^{L}(-1)^{k+1}\left(\cos\frac{k\pi}{m+1}-R\right)^{m-2}\sin^2\frac{k\pi}{m+1},$$

where

$$\cos\frac{(L+1)\pi}{m+1} \leqq R \leqq \cos\frac{L\pi}{m+1}.$$

For $\rho = 0$ it is clear that

$$g(R) = 2^m(m-1)(m+1)^{-1}\sum_{k=1}^{L}(-1)^{k+1}$$

$$\cdot\left(\cos\frac{k\pi}{m+1}-R\right)^{m-2}\sin^2\frac{k\pi}{m+1} \tag{3.6}$$

and hence

$$g(R;\rho) = (1-\rho^{2m+2})(1-\rho^2)^{-1}(1+\rho^2-2\rho R)^{-m}g(R). \tag{3.7}$$

4. Approximations. In a paper by T. W. Anderson and R. L. Anderson [3] in which the circular serial correlation coefficient is discussed for fitted trigonometric series for the mean, they have fitted the trigonometric series for semi-annual data to correct for variation of period two and get a quadratic form

$$q = X'CX \,/\, X'BX \tag{4.1}$$

for $n = 2m$.

They reduce q to the form

$$\sum_{j=1}^{2m-2} c_j y_j^2 \Big/ \sum_{j=1}^{2m-2} y_j^2, \tag{4.2}$$

where the c_j are identical with the a_j of the previous section.

Therefore the distribution in this particular case for $2m$ observations is exactly the same as that for the non-circular case of Durbin and Watson for $2m-2$ observations when $\rho = 0$. They also give the approximate distribution of their circular statistic as a beta distribution, and if we put $2m-2$ in place of $2m$ we get the approximate distribution density of R for $2m-2$ observations

when $\rho = 0$ to be

$$g(R) \sim K(1 - R^2)^p, \tag{4.3}$$

where $p = (m^2 + m)(m - 1)^{-1} - 3/2$.

5. Moments of a ratio. The previous work was based on the assumption that $\mathcal{E}x_i = \mu$ was known. However, if μ is unknown, one of the statistics used is the ratio of the mean square successive difference to the variance;

$$\eta = \frac{\delta^2}{s^2} \tag{5.1}$$

The distribution of η is for the present too difficult for explicit evaluation. The moments of η have been found by Williams [10] and much light shed on the distribution by Von Neumann [8]. However, the expression for the rth moment given by Williams is in terms of the rth derivative of a function.

Durbin and Watson [4] suggested a modified statistic in this case for $n = 2m$. Let

$$R = X'AX \,/\, X'BX \quad \text{or} \quad R = 4(m-1)\delta_0^2/[(2m-1)s^2], \tag{5.2}$$

where

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_1 \end{pmatrix},$$

$$A_1 = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & & & \\ & \ddots & \ddots & & \\ & & & 2 & -1 \\ & & & -1 & 1 \end{pmatrix},$$

with latent roots

$$a_j = 4 \sin^2 \frac{(m-j)\pi}{2m} = 4 \cos^2 \frac{j\pi}{2m},$$

and

$$B = \frac{1}{2m} \begin{bmatrix} 2m-1 & -1 & \cdots & -1 \\ -1 & & & \vdots \\ \vdots & & & \vdots \\ \vdots & & & -1 \\ -1 & \cdots & & 2m-1 \end{bmatrix}.$$

The distribution of R is given to be

$$P(R > R') = \sum_{k=1}^{L} (a_k - R')^{m-3/2} a_k^{-\frac{1}{2}} \prod_{\substack{j=1 \\ j \neq k}}^{m-1} (a_k - a_j)^{-1} \tag{5.3}$$

for $a_{k+1} \leqq R \leqq a_k$. This result is based on a result of Anderson's where in addition to double roots there is a single root which is less than all of the double roots. By simplification it can be reduced to

$$P(R > R') = 4m^{-1} \sum_{k=1}^{L} (-1)^{k+1} (a_k - R')^{m-3/2} \sin^2 \frac{k\pi}{2m} \cos \frac{k\pi}{2m} \tag{5.4}$$

for $a_{k+1} \leqq R \leqq a_k$. The moments of this ratio can be easily found for $r < 3m - 1$, since we already have evaluated the moments of the numerator in a previous paper (Geisser [5]) and the moments of the denominator are well known. When the successive observations are independent, the moments of this ratio are the ratio of the moments [9]. Therefore

$$\mathcal{E}R^r = \mu_r = \frac{2^{r+1}(2m + 2r - 2)!(2m^2 - m - r)}{[(2m + r)!(2m - 1)(2m + 1) \cdots (2m + 2r - 3)]}. \tag{5.5}$$

6. The distribution of the modified von Neumann ratio. If we consider the ratio

$$\eta_0 = \frac{2\delta_0^2}{s_0^2} = \frac{\sum_{\substack{i=1 \\ i \neq m}}^{2m-1} (x_{i+1} - x_i)^2}{\sum_{1}^{m} (x_i - \bar{x}_1)^2 + \sum_{m+1}^{2m} (x_i - \bar{x}_2)^2}, \tag{6.1}$$

i.e., twice the ratio of the modified mean square successive difference to the pooled variance, we are able to use the Double Root Result and find the distribution of η_0 for the non-null alternative given by T. W. Anderson [2] if we further consider the model to be made up of two independent sets and

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_1 \end{pmatrix}, \tag{6.2}$$

$$\Lambda_1 = \sigma^{-2} \begin{bmatrix} 1 + \rho^2 - \rho & -\rho & & & \\ -\rho & 1 + \rho^2 & & & \\ & & \ddots & & \\ & & & 1 + \rho^2 & -\rho \\ & & & -\rho & 1 + \rho^2 - \rho \end{bmatrix}. \tag{6.3}$$

As was shown previously [2], η_0 provides a uniformly most powerful one-sided test. By the double root result we get, letting $\lambda_j' = \sigma^2 \lambda_j$,

$$G(\eta_0;\rho) = 1 - 2m^{-1}[1+\rho^2-2\rho+\rho\eta_0]^{2-m}\prod_{j=1}^{m-1}\lambda_j' \cdot \sum_{k=1}^{L}(-1)^{k+1}(a_k-\eta_0)^{m-2}\lambda_k^{-1}\sin^2\frac{k\pi}{m} \tag{6.4}$$

and

$$G'(\eta_0;\rho) = g(\eta_0;\rho) = 2m^{-1}(m-2)[1+\rho^2-2\rho+\rho\eta_0]^{2-m}\prod_{j=1}^{m-1}\lambda_j' \cdot \sum_{k=1}^{L}(-1)^{k+1}(a_k-\eta_0)^{m-3}\sin^2\frac{k\pi}{m}, \tag{6.5}$$

where $a_{L+1} \leqq \eta_0 \leqq a_L$. For $\rho = 0$,

$$G(\eta_0) = 2(m-2)m^{-1}\sum_{k=1}^{L}(a_k-\eta_0)^{m-2}(-1)^{k+1}\sin^2\frac{k\pi}{m} \tag{6.6}$$

and

$$G'(\eta_0) = g(\eta_0) = 2(m-2)m^{-1}\sum_{k=1}^{L}(-1)^{k+1}(a_k-\eta_0)^{m-3}\sin^2\frac{k\pi}{m} \tag{6.7}$$

for $a_{L+1} \leqq \eta_0 \leqq a_L$. Hence

$$G'(\eta_0;\rho) = g(\eta_0;\rho) = (1+\rho^2-2\rho+\rho\eta_0)^{1-m}\left(\prod_{j=1}^{m-1}\lambda_j'\right)g(\eta_0); \tag{6.8}$$

and since

$$\prod_{j=1}^{m-1}\lambda_j' = \frac{(1-\rho)(1-\rho^{2m})}{(1+\rho)(1-\rho)^2} = \frac{1-\rho^{2m}}{1-\rho^2},$$

$$G'(\eta_0;\rho) = (1+\rho^2-2\rho+\rho\eta_0)^{1-m}(1-\rho^{2m})(1-\rho^2)^{-1}g(\eta_0). \tag{6.9}$$

It is also quite easy to find the moments $\mathcal{E}\eta_0^r$ when $\rho = 0$ since we have already given the moments of the numerator for $r < 3m-1$ and the moments of the denominator are well known. Hence for $r < 3m-1$

$$\mathcal{E}\eta_0^r = \frac{2^{1-r}(2m^2-m-r)(2m+2r-2)!(m-2)!}{(m+r-2)!(2m+r)!}. \tag{6.10}$$

7. Acknowledgements. I am grateful to Dr. Harold Hotelling for his guidance and to the referee for his helpful comments.

REFERENCES

[1] R. L. Anderson, "Distribution of the serial correlation coefficient," *Ann. Math. Stat.*, Vol. 13 (1942), pp. 1–13.

[2] T. W. Anderson, "On the theory of testing serial correlation," *Skand. Aktuarietids.*, Vol. 31 (1948), pp. 88–115.

[3] T. W. Anderson and R. L. Anderson, "Distribution of the circular serial correlation coefficient for residuals from a fitted Fourier series," *Ann. Math. Stat.*, Vol. 21 (1950), pp. 59–81.

[4] J. Durbin and G. S. Watson, "Exact tests of serial correlation using non-circular statistics," *Ann. Math. Stat.*, Vol. 22 (1951), pp. 446–451.

[5] S. Geisser, "The modified mean square successive difference and related statistics," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 819–824.

[6] J. Gurland, "Inversion formulae for the distribution of ratios," *Ann. Math. Stat.*, Vol. 19 (1948), pp. 228–237.

[7] W. G. Madow, "Note on the distribution of the serial correlation coefficient," *Ann. Math. Stat.*, Vol. 16 (1945), pp. 308–310.

[8] J. von Neumann, "Distribution of the ratio of the mean square successive difference to the variance," *Ann. Math. Stat.*, Vol. 12 (1941), pp. 367–395.

[9] J. von Neumann, "A further remark concerning the distribution of the ratio of the mean square successive difference to the variance," *Ann. Math. Stat.*, Vol 13 (1942), pp. 86–88.

[10] J. D. Williams, "Moments of the ratio of the mean square successive difference to the variance in samples from a normal universe," *Ann. Math. Stat.*, Vol. 12 (1941), pp. 239–241.

RESTRICTION AND SELECTION IN MULTINORMAL DISTRIBUTIONS[1]

By A. Clifford Cohen, Jr.

The University of Georgia

1. Summary. Maximum likelihood estimators of the parameters of a p-dimensional multinormal population are derived in this paper which are applicable when sample selection and observation is restricted with respect to x_1 but otherwise unrestricted with respect to $x_2, \cdots, x_p$. Restrictions imposed may consist of truncation, censoring, or a selection which results in full observation of all sample specimens with respect to x_1, but eliminates certain sample specimens from subsequent observation with respect to $x_2, \cdots, x_p$.

2. Introduction. Samples from a multidimensional universe are often obtained under circumstances such that observation in certain regions of the universe is restricted. For example, in studies of psychological traits, observation is often limited to individuals who have passed certain admission tests or who have been subjected to other screening processes. This situation likewise arises in connection with multivariate studies of physical characteristics in which specimens available for observation have previously undergone some type of sorting procedure. From such samples, it is often necessary to estimate the means, variances and correlation coefficients of the universe. Considering their most general aspects without limitation as to type of distribution, restricted or "screened" samples pose a broad class of estimation problems, some of which are quite involved. The present paper is limited to samples from a p-dimensional multinormal distribution with probability density function

$$(1) \quad f(x_1, x_2, \cdots, x_p) = (2\pi)^{-p/2} |\sigma^{ij}|^{1/2} \exp\left[-\frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \sigma^{ij}(x_i - m_i)(x_j - m_j) \right],$$

where the symmetric matrix $\|\sigma^{ij}\|$ of the quadratic form in the exponent is the inverse of the variance-covariance matrix $\|\sigma_{ij}\|$, and has the positive determinant $|\sigma^{ij}|$. Maximum likelihood estimators (estimates) for parameters of (1) are obtained from truncated, censored and selected samples, with x_1 designating the restricted variable; that is, the variable on which screening is based. Similar estimators obtained previously ([8], [10]) for restricted samples from a bivariate normal distribution, follow as a special case of results obtained here. Results obtained by Hotelling [12], Tukey [18], Pittman [16], and Chapman [5] guarantee that both the method of moments and the method of maximum likelihood lead to identical estimates in the case of truncated samples from multinormal distributions. Hence for truncated samples we might have employed the method of moments. However, we elected to use the method of maximum likelihood

Received November 26, 1954; revised June 30, 1956.

[1] Sponsored by the Office of Ordnance Research, U. S. Army.

as it permits a uniform treatment of all the various types of restricted samples under consideration and it introduces no unusual algebraic difficulties.

In practical applications such as in studies of psychological traits, the screening variable x_1 might actually be a composite score based on a battery of tests rather than the score achieved on a single test. However, in this paper, we limit our consideration to cases in which each of the component variables, including x_1, has a univariate normal marginal distribution. In some applications, one or more achievement scores might be involved which also are composite scores. For example, x_p could be such a score. Here again, however, the limitation of normality on marginal distributions holds.

Various aspects of some of the basic problems involved in the present study have previously been investigated by Karl Pearson [15], Aitken [1], Wilks [20], Birnbaum Paulson and Andrews [3], Votaw, Rafferty and Deemer [19], Campbell [4], Des Raj [17], and the author [8], [9], [10]. A more complete bibliography of related papers can be found in reference [7].

3. Estimating means, standard deviations, and correlation coefficients. For a random sample of n (fixed) measured observations $(x_{1\alpha}, x_{2\alpha}, \cdots, x_{p\alpha})$, $\alpha = 1, 2, \cdots n$, drawn from a population distributed according to (1), subject to a restriction on observation of variable x_1, the logarithm of the likelihood function is

$$L = -(np/2)\ln 2\pi + (n/2)\ln|\sigma^{ij}| - \tfrac{1}{2}\sum_i\sum_j\sum_\alpha \sigma^{ij}(x_{i\alpha} - m_i)(x_{j\alpha} - m_j) + \ln G(m_1, \sigma_{11}), \tag{2}$$

where $G(m_1, \sigma_{11})$ is a restriction function which depends upon the type of restriction imposed with respect to observation of x_1 by screening or acceptance criteria. When G is to be interpreted with full generality, it is not only a function of m_1 and σ_{11}, but also of x_1. By thus introducing G, much repetition in the derivation of estimators is avoided which otherwise would arise with the various selection criteria to be considered. Specific examples of G are given subsequently in this paper.

For an unrestricted sample, $G(m_1, \sigma_{11}) \equiv 1$, and maximum likelihood estimates of parameters m_i and σ^{ij} are obtained by equating to zero, the partial derivatives of L with respect to these parameters and solving the resulting system of equations. (Cf. for example Mood [14], pp. 186–188.) In the cases involving restricted or screened samples, we follow a similar procedure. However, in order to avoid certain complications which restrictions on x_1 introduce into derivatives with respect to σ^{ij}, we employ derivatives with respect to σ_{11} and ρ_{ij}. According to the notation employed here, $\sigma_{ij} = \sigma_i\sigma_j\rho_{ij}$ and $\sigma_{11} = \sigma_1^2$, where ρ_{ij} is the coefficient of correlation between x_i and x_j.

Considering first the means, we have

$$(3)\quad \text{(a)}\quad \frac{\partial L}{\partial m_1} = n\sum_{i=1}^{p}\sigma^{1i}C_{i0} + \frac{1}{G}\frac{\partial G}{\partial m_1},$$

$$\text{(b)}\quad \frac{\partial L}{\partial m_r} = n\sum_{i=1}^{p}\sigma^{ri}C_{i0}, \qquad r = 2, 3, \cdots, p,$$

where

$$(4)\quad C_{ij} = \sum_{\alpha}(x_{i\alpha} - m_i)(x_{j\alpha} - m_j)/n \qquad \text{and} \qquad C_{i0} = \sum_{\alpha}(x_{i\alpha} - m_i)/n.$$

On setting $\partial L/\partial m_r = 0$, $(r \geqq 2)$ and dividing by C_{10}, we obtain the system of $p - 1$ equations:

$$(5)\quad \sum_{i=1}^{p}\sigma^{ri}C_{i0}/C_{10} = 0, \qquad r = 2, 3, \cdots, p.$$

On solution, these yield the result

$$(6)\quad C_{i0}/C_{10} = A_{1i}/A_{11},$$

where A_{ij} is the cofactor of $|\sigma^{ij}|$. Since $\|\sigma^{ij}\| = \|\sigma_{ij}\|^{-1}$, then $\sigma_{ij} = A_{ij}/|\sigma^{ij}|$ and $A_{ij} = \sigma_{ij}|\sigma^{ij}|$. On substituting this result into (6), we have

$$(7)\quad C_{i0}/C_{10} = [\sigma_{1i}|\sigma^{ij}|]/\sigma_{11}|\sigma^{ij}|] = \sigma_{1i}/\sigma_{11}.$$

After making the further substitution $\sigma_{1i} = \sigma_1\sigma_i\rho_{1i}$, it follows from (7) that

$$(8)\quad C_{i0} = \rho_{1i}(\sigma_i/\sigma_1)C_{10}.$$

We turn now to the variances and to the correlation coefficients. Since $C_{ij} = C_{ji}$ and $\sigma^{ij} = \sigma^{ji}$, we need only the derivatives

$$(9)\quad \text{(a)}\quad \frac{\partial L}{\partial \sigma_{11}} = -\frac{n}{2\sigma_{11}}\left\{1 - \sum_{j=1}^{p}\sigma^{ij}C_{ij}\right\} + \frac{1}{G}\frac{\partial G}{\partial \sigma_{11}},$$

$$\text{(b)}\quad \frac{\partial L}{\partial \sigma_{ss}} = -\frac{n}{2\sigma_{ss}}\left\{1 - \sum_{j=1}^{p}\sigma^{sj}C_{sj}\right\}, \qquad s = 2, 3, \cdots, p,$$

$$\text{(c)}\quad \frac{\partial L}{\partial \rho_{rs}} = -n\sigma_r\sigma_s\left\{\sigma^{rs} - \sum_{i=1}^{p}C_{ii}\sigma^{ir}\sigma^{is}\right.$$

$$\left. - \sum_{i=1}^{p-1}\sum_{i<j}^{p}\sum_{j=2}C_{ij}[\sigma^{rs}\sigma^{ij} + \sigma^{ri}\sigma^{js}(1 - \delta_{ij}^{rs}) + \sigma^{si}\sigma^{jr}\delta_{ij}^{rs}]\right\},$$

$$r = 1, 2, \cdots, p - 1;\ s = 2, 3, \cdots, p;\ r < s,$$

where δ_{ij}^{rs} is a generalized form of Kronecker's delta such that it has the value 1 if $\sigma^{ri}\sigma^{js} = \sigma^{rs}\sigma^{ij}$, but otherwise it has the value zero.

We equate to zero, the $(p - 1)$ derivatives $\partial L/\partial\sigma_{ss}$ of (9b) and the $p(p - 1)/2$ derivatives $\partial L/\partial\rho_{rs}$ of (9c) to form a total of $(p - 1)(p + 2)/2$ equations that

are linear in $C_{ij}(i \leqq j)$. These we now write as

$$\sum_{i=1}^{p} \sigma^{is} C_{is} - 1 = 0, \qquad s = 2, 3, \cdots, p,$$

$$\sum_{i=1}^{p} C_{ii}\sigma^{ir}\sigma^{is} + \sum_{i=1}^{p-1}\sum_{i<j}^{p} C_{ij}[\sigma^{ri}\sigma^{sj} + \sigma^{ri}\sigma^{jr}(1 - \delta_{ij}^{rs}) + \sigma^{si}\sigma^{jr}\delta_{ij}^{rs}] - \sigma^{rs} = 0, \quad r = 1, 2, \cdots, p-1; s = 2, \cdots, p; r < s. \tag{10}$$

As a solution of this system of equations, we obtain C_{ij} in terms of C_{11} as

$$C_{ij} = \sigma_{ij} + (\sigma_{1i}\sigma_{1j}/\sigma_{11})(C_{11}/\sigma_{11} - 1), \qquad i \leqq j, \tag{11}$$

which can be verified by direct substitution back into the equations of (10). As a special case of (11), we have

$$C_{1i} = (\sigma_{1i}/\sigma_{11})C_{11}\,. \tag{11a}$$

Returning now to the definitions for C_{ij} and C_{i0} as given in (4), we can write

$$\begin{aligned} C_{ij} - C_{i0}C_{j0} &= \sum_{\alpha}(x_{i\alpha} - m_i)(x_{j\alpha} - m_j)/n \\ &\quad - [\sum_{\alpha}(x_{i\alpha} - m_i)/n][\sum_{\alpha}(x_{j\alpha} - m_j)/n] \\ &= [\sum_{\alpha} x_{i\alpha}x_{j\alpha}/n - m_j\bar{x}_i - m_i\bar{x}_j + m_im_j] \\ &\quad - [\bar{x}_i\bar{x}_j - m_i\bar{x}_j - m_j\bar{x}_i + m_im_j] \end{aligned}$$

and thus

$$C_{ij} - C_{i0}C_{j0} = \sum_{\alpha} x_{i\alpha}x_{j\alpha}/n - \bar{x}_i\bar{x}_j\,, \quad \alpha = 1, 2, \cdots, n, \tag{12}$$

where $\bar{x}_k = \sum_{\alpha} x_{k\alpha}/n$.

With restricted sample standard deviations written as $\bar{s}_i$ and restricted sample correlation coefficients written as $\bar{r}_{ij}$ where

$$\begin{aligned} \bar{s}_i &= \left[\sum_{\alpha=1}^{n} x_{i\alpha}^2/n - \left(\sum_{\alpha=1}^{n} x_{i\alpha}/n\right)^2\right]^{1/2} \\ \bar{r}_{ij} &= \left[n\sum_{\alpha=1}^{n} x_{i\alpha}x_{j\alpha} - \sum_{\alpha=1}^{n} x_{i\alpha}\sum_{\alpha=1}^{n} x_{j\alpha}\right] \Big/ n^2\bar{s}_i\bar{s}_j\,, \end{aligned} \tag{13}$$

Eq. (12) becomes

$$C_{ij} - C_{i0}C_{j0} = \bar{r}_{ij}\bar{s}_i\bar{s}_j\,. \tag{14}$$

Since $\sigma_{ij} = \sigma_i\sigma_j\rho_{ij}$, Eqs. (8) and (11) permit us to write

$$C_{ij} - C_{i0}C_{j0} = \sigma_i\sigma_j[\rho_{ij} - \lambda\rho_{1i}\rho_{1j}], \tag{15}$$

where

$$\lambda = 1 - \bar{s}_1^2/\sigma_1^2\,. \tag{16}$$

Equating the right side of (14) to the right side of (15), we have

$$\bar{r}_{ij}\bar{s}_i\bar{s}_j = \sigma_i\sigma_j[\rho_{ij} - \lambda\rho_{1i}\rho_{1j}]. \tag{17}$$

We let $i = j$, and since $\bar{r}_{ii} \equiv 1$ and $\rho_{ii} \equiv 1$, it follows from (17) that

$$\sigma_j^2 = \bar{s}_j^2/(1 - \rho_{1j}^2\lambda), \qquad j = 2, 3, \cdots, p. \tag{18}$$

Let $i = 1$, eliminate σ_j between (17) and (18), and we have

$$\rho_{1j} = \bar{r}_{1j}/\sqrt{1 - \lambda(1 - \bar{r}_{1j}^2)}, \qquad j = 2, 3, \cdots, p. \tag{19}$$

Use (18) to write first σ_i and then σ_j. Substitute these results into (17) and simplify to obtain

$$\rho_{ij} = \bar{r}_{ij}\sqrt{(1 - \rho_{1i}^2\lambda)(1 - \rho_{1j}^2\lambda)} + \rho_{1i}\,\rho_{1j}\,\lambda, \tag{20}$$

with $i, j = 2, 3, \cdots, p$, $i < j$.

Estimates $\hat{m}_1$ and $\hat{\sigma}_1$ are yet to be determined, but Eqs. (8), (18), (19) and (20) enable us to express estimators for the remaining parameters of (1) in terms of these two as

$$\begin{aligned}
\hat{m}_j &= \bar{x}_j - \bar{r}_{1j}(\bar{s}_j/\bar{s}_1)(\bar{x}_1 - \hat{m}_1), \\
\hat{\sigma}_j &= \bar{s}_j\sqrt{\frac{1 - \hat{\lambda}(1 - \bar{r}_{1j}^2)}{1 - \hat{\lambda}}}, \\
\hat{\rho}_{ij} &= \frac{\bar{r}_{ij} - \hat{\lambda}(\bar{r}_{ij} - \bar{r}_{ij}\bar{r}_{ij})}{\sqrt{[1 - \hat{\lambda}(1 - \bar{r}_{1i}^2)][1 - \hat{\lambda}(1 - \bar{r}_{1j}^2)]}},
\end{aligned} \tag{21}$$

with $i = 1, 2, \cdots, p - 1$, $j = 2, 3, \cdots, p$, $i < j$, and $\hat{\lambda} = 1 - \bar{s}_1^2/\delta_1^2$. Since by definition, $\bar{r}_{ii} \equiv 1$, the last equation of (21), in agreement with (19), simplifies to

$$\hat{\rho}_{1j} = \bar{r}_{1j}/\sqrt{1 - \hat{\lambda}(1 - \bar{r}_{1j}^2)}, \tag{22}$$

when $i = 1$. Here and throughout this paper, the maximum likelihood symbol ($\hat{}$) serves to distinguish estimates from the parameters estimated.

Although the derivations were somewhat more laborious, the above results were given earlier in [9]. Estimators for restricted samples from a bivariate normal population as given in [8] and [10] now follow as a special case of (21) and (22) with $j = p = 2$.

To estimate m_1 and σ_1, we substitute (7) into (3a) and (11a) into (9a), equate to zero, and thereby obtain

$$\begin{aligned}
\frac{\sigma_{11}}{n}\frac{\partial L}{\partial m_1} &= C_{10}\sum_1^p \sigma^{1i}\sigma_{1i} + \frac{\sigma_{11}}{nG}\frac{\partial G}{\partial m_1} = 0, \\
\frac{2\sigma_{11}}{n}\frac{\partial L}{\partial \sigma_{11}} &= \frac{C_{11}}{\sigma_{11}}\sum_1^p \sigma^{1i}\sigma_{1i} - 1 + \frac{2\sigma_{11}}{nG}\frac{\partial G}{\partial \sigma_{11}} = 0.
\end{aligned} \tag{23}$$

Since $\sum_i \sigma^{im}\sigma_{mj} = \delta_{ij}$ (cf., for example, [14], p. 179), where $\delta_{ij} = 1$, if $i = j$, and $\delta_{ij} = 0$, if $i \neq j$, it follows that $\sum_1^p \sigma^{1i}\sigma_{1i} = 1$, and with the defining rela-

tions for C_{10} and C_{11} as given by (4), this result enables us to write

$$\sum_{\alpha=1}^{n} (x_{1\alpha} - m_1)/n + \frac{\sigma_{11}}{nG}\frac{\partial G}{\partial m_1} = 0,$$

$$\sum_{\alpha=1}^{n} (x_{1\alpha} - m_1)^2/n - \sigma_{11}\left[1 - \frac{2\sigma_{11}}{nG}\frac{\partial G}{\partial \sigma_{11}}\right] = 0. \tag{24}$$

The required estimates $\hat{m}_1$ and $\hat{\sigma}_1$ are the values found on solving this pair of equations. The restriction function, $G(m_1, \sigma_{11})$ which depends upon the nature of the restrictions imposed on x_1 must be specified before Eqs. (24) are completely determined, but it is to be noted that regardless of G, they involve only the x_1-marginal distribution and are independent of the remaining variables.

Truncated and censored samples. When samples under consideration have been truncated or censored with respect to x_1, estimating Eqs. (24) reduce to forms identical with those obtained previously in reference [7] in connection with various types of truncated and censored samples from univariate normal populations. They can be solved as therein described for the univariate cases. For example, when x_1 is *singly truncated* on the left at a fixed terminal x_{10}, then $G(m_1, \sigma_{11}) = [I_0(\xi)]^{-n}$, where $I_0(\xi) = \int_\xi^\infty \varphi(t)\,dt$, $\varphi(t) = [(2\pi)^{1/2}]^{-1} \exp(-t^2/2)$, and $\xi = (x_{10} - m_1)/\sigma_1$. In this case, estimating Eqs. (24) reduce to

$$\text{(a)} \quad \frac{1}{\hat{Z} - \hat{\xi}}\left[\frac{1}{\hat{Z} - \hat{\xi}} - \hat{\xi}\right] = \frac{n\sum_{\alpha=1}^{n} (x_{1\alpha} - x_{10})^2}{\left[\sum_{\alpha=1}^{n} (x_{1\alpha} - x_{10})\right]^2},$$

$$\text{(b)} \quad \hat{\sigma}_1 = \sum_{\alpha=1}^{n} (x_{1\alpha} - x_{10})/n(\hat{Z} - \hat{\xi}), \tag{25}$$

$$\text{(c)} \quad m_1 = x_{10} - \hat{\sigma}_1\hat{\xi},$$

where

$$Z(\xi) = \varphi(\xi)/I_0(\xi) = \exp(-\xi^2/2) \Big/ \int_\xi^\infty \exp(-t^2/2)\,dt. \tag{26}$$

Equation (25a) can be solved for $\hat{\xi}$, so that $\hat{\sigma}_1$ and $\hat{m}_1$ follow in turn from (25b) and (25c). For further details, reference is again made to [7]. Whenever m_1 and σ_1 are known a priori, the remaining parameters can be estimated from (21) with $\hat{\lambda} = 1 - s_1^2/\hat{\sigma}_1^2$ replaced by $\lambda = 1 - s_1^2/\sigma_1^2$ and $\hat{m}_1$ replaced by the known value of m_1.

Selected samples. When sampling procedure is such that a total of N unrestricted observations are made with respect to x_1 although as a result of selection or screening, there may be only $n(< N)$ observations of $x_2, \cdots, x_p$, then

$$G(m_1, \sigma_{11}) = (\sqrt{2\pi\sigma_{11}})^{n-N} \exp\left[-\sum_{1}^{N-n} (x_{1\alpha} - m_1)^2/2\sigma_{11}\right],$$

and Eqs. (24) lead to the familiar estimates

$$(27) \qquad \hat{m}_1 = \sum_1^N x_{1a}/N = \bar{x}_1, \qquad \hat{\sigma}_1^2 = \sum_1^N (x_{1a} - \bar{x}_1)^2/N.$$

Regardless of how the selection which determines subsequent observation with respect to $x_2, \cdots, x_p$ is made, $\hat{m}_1$ and $\hat{\sigma}_1$ are given by (27) while the other estimates are given by (21) where $\bar{x}_j$, $\bar{s}_j$, and $\bar{r}_{ij}$ are computed from observations of the n "selected" members of the sample.

Unrestricted samples. When no sample restrictions are imposed, and no selection is made, then not only is $G \equiv 1$, but $N = n$, $\lambda = 0$, and the required estmates follow from (27) and (21) as

$$(28) \qquad \hat{m}_j = \bar{x}_j, \qquad \hat{\sigma}_j = s_j, \qquad \hat{\rho}_{ij} = r_{ij}, \quad i, j = 1, 2, \cdots p,$$

which, as already mentioned, are well known for this case. The bars ($\bar{}$) are omitted over r_{ij} and s_j in (28) since here the computations are based on the complete rather than a restricted sample.

4. Reliability of estimates. Asymptotic variances and covariances of estimates given in the preceding section can, of course, be obtained from the likelihood information matrices with elements which are expected values of the second partial derivatives of the likelihood function L. These variances and covariances are of the order of $1/n$, but exact expressions for them are too unwieldy to be of much practical value. For parameters of the restricted variable, in this case x_1, asymptotic variances and covariances given in [7] for truncated and censored samples from univariate normal distributions are applicable when restrictions are of these types. When a selection based on x_1 is made which does not restrict observation of x_1 itself, then complete sample variances

$$V(\hat{m}_1) = \sigma_1^2/N,$$
$$V(\hat{\sigma}_1) = \sigma_1^2/2N,$$

are applicable as are various exact small sample results based on the x_1 marginal distribution. If the restrictions involved have not been unduly severe, that is, if only minor portions of the tails of the x_1 distribution have been affected, then asymptotic variances and covariances for complete (unrestricted) samples from a multinormal distribution will afford reasonably satisfactory approximations to the desired values. (Cf. Kendall [13], Vol. 11, third edition, p. 38.)

5. Practical applications. The practical application of estimators obtained in this paper is illustrated with a sample given by Baten [2], and attributed by him to H. C. Carver. The basic sample consists of weight, height, shoulder, chest, waist, and hip measurements on 119 individuals. We designate these variates in the order listed as x_1, x_2, x_3, x_4, x_5, and x_6 respectively. Baten's data include 120 sets of measurements, but it was necessary to eliminate the last one because of a typographical error. As given by Baten, the sample was considered to be complete, but for purposes of the illustrations here, it is arbitrarily truncated with respect to weight (x_1) at 119.5 pounds. Thereby eleven sets of measurements

are eliminated. Estimates of the population parameters are then computed considering the sample as truncated with $n = 108$, and as censored with $n = 108$ and $n_1 = 11$. A complete summary of estimates calculated for each of these two cases is included in Table 2 along with corresponding estimates computed from the complete sample. As can be observed from this table, estimates based on the truncated and censored samples are in close agreement with those computed from the complete sample. The computing procedures employed are illustrated below.

Truncated sample—number missing observations unknown. For this case, the sample data are summarized in Table 1.

To estimate parameters of x_1, we may follow the procedure described in [7] and first solve equation (25a), which for this example is

$$\frac{1}{\hat{Z} - \hat{\xi}}\left[\frac{1}{\hat{Z} - \hat{\xi}} - \hat{\xi}\right] = \frac{108(64{,}169.00)}{(2{,}301.0)^2} = 1.308928.$$

Thereby, we obtain $\hat{\xi} = -1.379$, and from (25b) we computed $\hat{\sigma}_1 = 13.7697$, and from (25c) $\hat{m}_1 = 119.5 - (13.7697)(-1.379) = 138.4884$. Tables [11] were employed to reduce the computing effort which otherwise would have been required.

From Eq. (16), we compute $\hat{\lambda} = 1 - \bar{s}_1^2/\hat{\sigma}_1^2 = 1 - (11.8419/13.7697)^2 = 0.2604$, and the remaining estimates are obtained from Eqs. (21). For illustration, specimen computations are given below.

$$\hat{m}_2 = 67.9241 - 0.4701(2.4008/11.8419)(21.3056 - 18.9884) = 67.7033,$$

$$\hat{\sigma}_2 = 2.4008\sqrt{\frac{1 - 0.2604(1 - 0.4701^2)}{1 - 0.2604}} = 2.4923,$$

$$\hat{\rho}_{12} = 0.4701/\sqrt{1 - (0.2604)(1 - 0.4701^2)} = 0.5265,$$

$$\hat{\rho}_{23} = \frac{0.2361 - 0.2604[0.2361 - (0.4701)(0.4326)]}{\sqrt{[1 - 0.2604(1 - 0.4701^2)]\,[1 - 0.2604(1 - 0.4326^2)]}} = 0.2872.$$

TABLE 1

Summary of Sample Data

$n = 108$	Truncation at $x_1 = 119.5$ lbs.		
$\bar{x}_1 = 140.8056$	$\bar{s}_1 = 11.8419$	$r_{12} = 0.4701$	$r_{25} = -0.1389$
	$\bar{s}_2 = 2.4008$	$r_{13} = 0.4326$	$r_{26} = 0.3019$
$\bar{x}_2 = 67.9241$	$\bar{s}_3 = 0.7103$	$r_{14} = 0.6501$	$r_{34} = 0.5904$
$\bar{x}_3 = 16.4500$	$\bar{s}_4 = 1.5373$	$r_{15} = 0.4415$	$r_{35} = 0.1852$
$\bar{x}_4 = 35.4537$	$\bar{s}_5 = 1.6375$	$r_{16} = 0.7873$	$r_{36} = 0.4059$
$\bar{x}_5 = 28.1574$	$\bar{s}_6 = 1.3746$	$r_{23} = 0.2361$	$r_{45} = 0.4931$
$\bar{x}_6 = 35.5898$		$r_{24} = 0.1194$	$r_{46} = 0.5491$
			$r_{56} = 0.4310$
$\sum_1^n (x_{1\alpha} - x_{10}) = 2301.0$		$\sum_1^n (x_{1\alpha} - x_{10})^2 = 64169.00$	

TABLE 2

Summary of Estimates

Parameters	Estimates Based on Complete Sample	Estimates Based on Restricted Sample	
		Truncated	Censored
		Number Missing Observations Unknown	Number Missing Observations Known
ξ		−1.379	−1.342
m_1	138.2353	138.4884	138.2382
m_2	67.6664	67.7033	67.6794
m_3	16.3672	16.3899	16.3834
m_4	35.1899	35.2581	35.2370
m_5	27.9252	28.0159	28.0007
m_6	35.3513	35.3780	35.3549
σ_1	13.9421	13.7697	13.9629
σ_2	2.5330	2.4923	2.5021
σ_3	0.7417	0.7333	0.7358
σ_4	1.7280	1.6477	1.6557
σ_5	1.7857	1.6927	1.6987
σ_6	1.5235	1.5172	1.5318
ρ_{12}	0.5239	0.5265	0.5318
ρ_{13}	0.5446	0.4872	0.4924
ρ_{14}	0.7339	0.7053	0.7037
ρ_{15}	0.5566	0.4966	0.5018
ρ_{16}	0.8369	0.8294	0.8330
ρ_{23}	0.2996	0.2872	0.2992
ρ_{24}	0.2406	0.2040	0.2114
ρ_{25}	−0.0120	−0.0613	−0.0536
ρ_{26}	0.3732	0.3772	0.3842
ρ_{34}	0.6193	0.6229	0.6265
ρ_{35}	0.3193	0.2365	0.2416
ρ_{36}	0.4908	0.4615	0.4667
ρ_{45}	0.5943	0.5362	0.5404
ρ_{46}	0.6569	0.6166	0.6220
ρ_{56}	0.5344	0.4849	0.4902
λ		0.2604	0.2806
Sample size	$n = 119$	$n = 108$	$n = 108$ $n_1 = 11$

Censored sample—number of missing (unmeasured) observations known. The sample data remain unchanged from the previous case except for the additional information that $n_1 = 11$. To estimate parameters of x_1, we determine ξ by

solving

$$\frac{1}{\hat{Y}-\xi}\left[\frac{1}{\hat{Y}-\xi}-\xi\right]=\frac{n\sum_{\alpha=1}^{n}(x_{1\alpha}-x_{10})^2}{\left[\sum_{\alpha=1}^{n}(x_{1\alpha}-x_{10})\right]^2}=1.308928,$$

where

$$Y(\xi)=\frac{n_1}{n}\left[\exp(-\xi^2/2)/(\sqrt{2\pi}-\int_{\xi}^{\infty}\exp(-t^2/2)\,dt)\right]=\frac{n_1}{n}Z(-\xi),$$

in the same manner as for the truncated case, and this time find $\hat{\xi}=-1.342$. Subsequently we compute $\hat{\sigma}_1=\sum_1^n(x_{1\alpha}-x_{10})/n(\hat{Y}-\xi)$, $=13.9629$. We then calculate $\hat{m}_1=119.5-(13.9629)(-1.342)=138.2382$. Using (16), we have $\hat{\lambda}=1-(11.8419/13.9629)^2=0.2806$. For further details, reference is again made to [7]. With $\hat{m}_1$ and $\hat{\sigma}_1$ thus determined, these values along with the original sample data are substituted into (21) to obtain estimates of the remaining parameters.

Although not complete in all details, the above calculations serve to indicate the general manner in which results of this paper are applicable in practical problems. To a certain extent, they also serve to indicate the degree of agreement to be expected among corresponding estimates based on truncated, censored and complete (unrestricted) samples.

REFERENCES

[1] A. C. AITKEN, "Note on selection from a multivariate normal population," *Proc. Math. Soc.*, Vol. 4 (1934), pp. 106–10.

[2] W. D. BATEN, *Mathematical Statistics*, John Wiley & Sons (1938), p. 193.

[3] Z. W. BIRNBAUM, E. PAULSON, AND F. C. ANDREWS, "On the effect of selection performed on some coordinates of a multi-dimensional population," *Psychometrika*, Vol. 15 (1950), pp. 191–204.

[4] FRANCIS L. CAMPBELL, "A study of truncated bivariate normal distributions," Doctoral Dissertation, University of Michigan (June, 1945).

[5] DOUGLAS G. CHAPMAN, "Sufficient statistics for selected distributions," *University of Washington Publication in Mathematics*, Vol. 3 (1952), pp. 59–64.

[6] A. C. COHEN, JR., "On estimating the mean and standard deviation of truncated normal distributions," *J. Amer. Stat. Assn.*, Vol. 44 (1949), pp. 518–25.

[7] A. C. COHEN, JR., "Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples," *Ann. Math. Stat.*, Vol. 21 (1950), pp. 557–69.

[8] A. C. COHEN, JR., "Estimation in truncated bivariate normal distributions," University of Georgia, Mathematical Technical Report No. 2, Contract DA-01-009-ORD-288, (June, 1953).

[9] A. C. COHEN, JR., "Estimation in truncated multivariate normal distributions," University of Georgia Mathematical Technical Report No. 3, Contract DA-01-009-ORD-288 (August, 1953).

[10] A. C. COHEN, JR., "Restriction and selection in samples from bivariate normal distributions," *J. Amer. Stat. Assn.*, Vol. 50 (1955), pp. 884–93.

[11] A. C. Cohen, Jr., and John Woodward, "Tables of Pearson-Lee-Fisher functions of singly truncated normal distributions," *Biometrics*, Vol. 9 (1953), pp. 489–97.

[12] Harold Hotelling, "Fitting generalized truncated normal distributions," Abstracts of Madison meeting, *Ann. Math. Stat.*, Vol. 19 (1948), p. 596.

[13] Maurice G. Kendall, *The Advanced Theory of Statistics*, 3d ed., Vol. 2, Charles Griffin and Co., Ltd., London, 1951, pp. 37–38.

[14] A. M. Mood, *Introduction to the Theory of Statistics*, McGraw Hill Book Co., 1950, pp. 165–191.

[15] Karl Pearson, "On the influence of natural selection on the variability and correlation of organs," *Philos. Trans. Roy. Soc. London*, Ser. A, Vol. 200 (1903), pp. 1–66.

[16] E. J. G. Pitman, "Sufficient statistics and intrinsic accuracy," *Proc. Cambridge Philos. Soc.*, Vol. 32 (1936), p. 567.

[17] Des Raj, "On estimating the parameters of bivariate normal populations from doubly and singly, linearly truncated samples," *Sankhya*, Vol. 12 (1953), pp. 277–90.

[18] John W. Tukey, "Sufficiency, truncation and selection," *Ann. Math. Stat.*, Vol. 20 (1949), pp. 309–11.

[19] D. F. Votaw, Jr., J. A. Rafferty, and W. L. Deemer, "Estimation of parameters in a truncated trivariate normal distribution," *Psychometrika*, Vol. 15 (1950), pp. 339–47.

[20] S. S. Wilks, "On estimates from fragmentary data," *Ann. Math. Stat.* Vol. 3 (1932), pp. 163–96.

COMPONENTS OF VARIANCE ANALYSIS FOR PROPORTIONAL FREQUENCIES

By J. D. Bankier and R. E. Walpole

McMaster University and Virginia Polytechnic Institute

1. Summary. With the exception of papers by G. W. Snedecor, G. M. Cox, and H. F. Smith ([8], [9], [10]), there seems to be little about proportional frequencies in the literature. In this paper we consider two-way crossed classifications and two-way nested classifications. The expected values of the sums of squares are obtained in a form which is applicable to a variety of components of variance models. The tests of several hypotheses are considered.

2. The Type I model for two-way crossed classifications. We consider an experiment in which p treatments are applied to q blocks. The ith treatment is applied to the jth block n_{ij} times. The n_{ij}'s having been displayed in a matrix with n_{ij} in the ith row and jth column, we assume that the n_{ij}'s in a given row are proportional to the n_{ij}'s in any other row. This implies that

$$n_{ij} = \frac{n_{i.}\, n_{.j}}{N}, \tag{1}$$

where

$$n_{i.} = \sum_{j=1}^{q} n_{ij}, \qquad n_{.j} = \sum_{i=1}^{p} n_{ij}, N = \sum_{i=1}^{p} n_{i.}.$$

Consider the model

$$Y_{ijk_{ij}} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk_{ij}}, \quad i = 1, 2, \cdots, p; j = 1, 2, \cdots, q; \tag{2}$$

$k_{ij} = 1, 2, \cdots, n_{ij}$, where the $\epsilon_{ijk_{ij}}$'s are NID $(0, \sigma^2)$ and the parameters are subject to the conditions

$$\sum_{i=1}^{p} n_{i.}\tau_i = \sum_{j=1}^{q} n_{.j}\beta_j = \sum_{i=1}^{p} n_{i.}(\tau\beta)_{ij} = \sum_{j=1}^{q} n_{.j}(\tau\beta)_{ij} = 0. \tag{3}$$

If we denote $E(Y_{ijk_{ij}})$ by ξ_{ij}, the above conditions are equivalent to defining

$$\mu = \bar{\xi}_{..}, \qquad \tau_i = \bar{\xi}_{i.} - \bar{\xi}_{..}, \qquad \beta_j = \bar{\xi}_{.j} - \bar{\xi}_{..}, \qquad (\tau\beta)_{ij} = \bar{\xi}_{ij} - \bar{\xi}_{i.} - \bar{\xi}_{.j} + \bar{\xi}_{..},$$

where

$$\bar{\xi}_{i.} = \frac{1}{n_{i.}} \sum_{j=1}^{q} n_{ij}\xi_{ij}, \qquad \bar{\xi}_{.j} = \frac{1}{n_{.j}} \sum_{i=1}^{p} n_{ij}\xi_{ij}, \qquad \bar{\xi}_{..} = \frac{1}{N} \sum_{i,j}^{p,q} n_{ij}\xi_{ij}.$$

A more realistic model, such as is considered by Anderson and Bancroft [1], will be studied in a later section.

We now rewrite Eqs. (2) in a form where the theory given by Anderson and

Received August 14, 1956, revised February 25, 1957.

Bancroft [1] may be applied. They may be put in the form

$$(4)\qquad Y_{ijk_{ij}} = \mu + \sum_{i'=1}^{p} U_{i'i}\tau_{i'} + \sum_{j'=1}^{q} V_{j'j}\beta_{j'} + \sum_{i',j'}^{p,q} W_{i'j'ij}(\tau\beta)_{i'j'} + \epsilon_{ijk_{ij}},$$

where

$$U_{i'i} = \delta_{i'i}, \qquad V_{j'j} = \delta_{j'j}, \qquad W_{i'j'ij} = \delta_{i'i}\delta_{j'j},$$

δ_{ij} being the Kronecker δ. If we order the $Y_{ijk_{ij}}$, calling them $Y_\alpha(\alpha = 1, 2, \cdots, N)$, we may write Eqs. (4) in the vector form

$$(5)\qquad Y = \mu + \sum_{i=1}^{p} U_i\tau_i + \sum_{j=1}^{q} V_j\beta_j + \sum_{i,j}^{p,q} W_{ij}(\tau\beta)_{ij} + \epsilon.$$

Denoting the elements of the vector U_i by $U_{i\alpha}$, we define

$$\overline{U}_i = \frac{1}{N}\sum_{\alpha=1}^{N} U_{i\alpha} = \frac{n_{i.}}{N}, \qquad u_{i\alpha} = U_{i\alpha} - \overline{U}_i$$

so that

$$0 = \sum_{\alpha=1}^{N} u_{i\alpha} = \sum_{i'=1}^{p} n_{i'.}u_{i'i'}.$$

Similarly,

$$\overline{V}_j = \frac{n_{.j}}{N}, \qquad \overline{W}_{ij} = \frac{n_{ij}}{N}, \qquad \sum_{j'=1}^{q} n_{.j'}v_{jj'} = \sum_{i',j'}^{p,q} n_{i'j'}w_{iji'j'} = 0.$$

Changing our notation, we denote by $\overline{U}_i$, $\overline{V}_j$, $\overline{W}_{ij}$, the vectors $\overline{U}_iI$, $\overline{V}_jI$, $\overline{W}_{ij}I$, where I is a column vector all of those N elements are equal to unity. Then, if we set

$$u_i = U_i - \overline{U}_i, \qquad v_j = V_j - \overline{V}_j, \qquad w_{ij} = W_{ij} - \overline{W}_{ij},$$

we may write Eq. (5) in the form

$$(6)\qquad Y = \mu + \sum_{i=1}^{p} u_i\tau_i + \sum_{j=1}^{q} v_j\beta_j + \sum_{i,j}^{p,q} w_{ij}(\tau\beta)_{ij} + \epsilon.$$

It is necessary that the u_i's, v_j's, and w_{ij}'s form a linearly independent set of vectors. Since this is not the case, we use the conditions (3) to eliminate τ_p, β_q, $(\tau\beta)_{iq}(i = 1, 2, \cdots, p)$ and $(\tau\beta)_{pj}(j = 1, 2, \cdots, q-1)$, obtaining

$$(7)\qquad \begin{aligned} Y = \mu &+ \sum_{i=1}^{p-1}\left(u_i - \frac{n_{i.}}{n_{p.}}u_p\right)\tau_i + \sum_{j=1}^{q}\left(v_j - \frac{n_{.j}}{n_{.q}}v_q\right)\beta_j \\ &+ \sum_{i=1}^{p-1}\sum_{j=1}^{q-1} n_{ij}\left(\frac{w_{ij}}{n_{ij}} - \frac{w_{iq}}{n_{iq}} - \frac{w_{pj}}{n_{pj}} + \frac{w_{pq}}{n_{pq}}\right)(\tau\beta)_{ij} + \epsilon. \end{aligned}$$

We note that

$$u_i - \frac{n_{i.}}{n_{p.}}u_p = U_i - \frac{n_{i.}}{n_{p.}}U_p$$

and a similar statement may be made about the coefficient vectors of β_j and $(\tau\beta)_{ij}$. Making use of the relations

$$U_i.U_{i'} = n_{i.}\delta_{ii'}, \qquad U_i.V_j = n_{ij}, \qquad U_i.W_{i'j} = n_{ij}\delta_{ii'}, \qquad V_j.V_{j'} = n_{.j}\delta_{jj'},$$

$$V_j.W_{ij'} = n_{ij'}\delta_{jj'}, \qquad W_{ij}.W_{i'j'} = n_{ij}\delta_{ii'}\delta_{jj'},$$

it may be proved that the coefficient vectors of the τ's, β_j's, $(\tau\beta)_{ij}$'s form three sets of linearly independent vectors and a vector from any set is orthogonal to the vectors of the other two sets. Thus, when the three sets are combined, they form a set of linearly independent vectors.

We shall be interested in testing three hypotheses

$$H_1: (\tau\beta)_{ij} = 0, \qquad i = 1, 2, \cdots, p-1; j = 1, 2, \cdots, q-1,$$

$$H_2: \tau_i = 0, \qquad i = 1, 2, \cdots, p-1,$$

$$H_3: \beta_j = 0, \qquad j = 1, 2, \cdots, q-1.$$

The restrictions (3) imply that not only the parameters in a given hypothesis are zero but also all other parameters of the same kind. To test H_1, we first compute

$$SSE = \sum_{i,j,k_{ij}} [Y_{ijk_{ij}} - m - t_i - b_j - (tb)_{ij}]^2,$$

where m, t_i, b_j and $(tb)_{ij}$ are the least squares estimates of μ, τ_i, β_j, and $(\tau\beta)_{ij}$, respectively, and SSE is the minimized value of the residual sum of squares. Next, we compute SSE_1, the corresponding minimum obtained under the assumption that H_1 holds. Then

$$R = \sum_{\alpha=1}^{N} y_\alpha^2 - SSE,$$

where $y_\alpha = Y_\alpha - \bar{Y}$, is the reduction in the sum of squares when all the parameters are used while

$$R_1 = \sum_{\alpha=1}^{N} y_\alpha^2 - SSE_1$$

is the reduction due to the parameters left when H_1 is true. The additional reduction in the sum of squares due to the $(\tau\beta)_{ij}$'s is

$$SS(TB) = R - R_1 = SSE_1 - SSE.$$

In the same way, SSE_2 and SSE_3 denote the minima obtained subject to H_2 and H_3, respectively, and the reductions in the sum of squares due to the τ_i's and the β_j's are

$$SST = SSE_2 - SSE \quad \text{and} \quad SSB = SSE_3 - SSE,$$

respectively. Anderson and Bancroft [1] show that

$$\sum_{\alpha=1}^{N} y_\alpha^2 = SST + SSB + SS(TB) + SSE,$$

and that, subject to the corresponding hypotheses, SST, SSB, $SS(TB)$ and SSE are independently distributed as $\chi^2\sigma^2$ with $p-1$, $q-1$, $(p-1)(q-1)$, and $N-pq$ degrees of freedom. The hypotheses H_1, H_2, and H_3 are tested by the statistics

$$F_1 = \frac{MS(TB)}{MSE}, \qquad F_2 = \frac{MST}{MSE}, \qquad F_3 = \frac{MSB}{MSE},$$

respectively, where MSE, for example, is SSE divided by the corresponding number of degrees of freedom.

3. The sums of squares. The following theorem, a slight generalization of one stated by Mann [5], will be used in computing the sums of squares.

THEOREM A. *If*

$$E(Y) = \mu I + \sum_{k=1}^{p} X_k \tau_k$$

and

(1) $I = \sum_{k=1}^{s} X_k, \qquad s \leqq p,$

(2) $X_1, X_2, \cdots, X_s$ *form a mutually orthogonal set of vectors,*

(3) $\sum_{k=1}^{s} n_k \tau_k = 0, \qquad \sum_{k=1}^{s} n_k \neq 0,$

(4) *any number of other conditions hold for* $\tau_{s+1}, \tau_{s+2}, \cdots, \tau_p$, *such that the method of Lagrange multipliers may be used,*

then condition (3) *may be ignored in the minimizing of*

$$SSE = \left(Y - \mu I - \sum_{k=1}^{p} X_k \tau_k\right)^2.$$

Our estimates of μ, τ_i, β_j, $(\tau\beta)_{ij}$ are m, t_i, b_j, $(tb)_{ij}$, respectively, where these values minimize SSE subject to the conditions (3). By Theorem A, we may ignore the conditions on the τ_i's and the β_j's. The conditions on the $(\tau\beta)_{ij}$'s will have to be considered in the computation of SSE_2 and SSE_3 but in the computation of SSE they can be avoided by expressing SSE in a different form. We have

$$SSE = \sum_{i,j,k_{ij}} (Y_{ijk_{ij}} - \xi_{ij})^2$$

Taking partial derivatives, we find our estimate of ξ_{ij} is $\hat{\xi}_{ij} = \bar{Y}_{ij.}$, where this notation indicates an average over the missing subscript. Then, by the invariance property of such estimators,

$$m = \bar{Y}_{...}, \qquad t_i = \bar{Y}_{i..} - \bar{Y}_{...}, \qquad bj = \bar{Y}_{.j.} - \bar{Y}_{...},$$

$$(tb)_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{..},$$

and

$$SSE = \sum_{i,j,k_{ij}} (Y_{ijk_{ij}} - \bar{Y}_{ij.})^2 = \sum_{i,j,k_{ij}} Y^2_{ijk_{ij}} - \sum_{i,j}^{p,q} \frac{Y^2_{ij.}}{n_{ij}}$$

where $Y_{ij.}$ is the sum of all the observations on the ith treatment in the jth block.

In obtaining SSE_1, all of the conditions (3) may be ignored, but, to determine SSE_2 and SSE_3, the method of Lagrange multipliers must be used. As a result of these calculations, we find that

$$SS(TB) = \sum_{i,j}^{p,q} n_{ij}(\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2,$$

$$SST = \sum_{i=1}^{p} n_{i.}(\bar{Y}_{i..} - \bar{Y}_{...})^2 = \sum_{i=1}^{p} \frac{Y_{i..}^2}{n_{i.}} - \frac{Y_{...}^2}{N},$$

and

$$SSB = \sum_{j=1}^{q} n_{.j}(\bar{Y}_{.j.} - \bar{Y}_{...})^2 = \sum_{j=1}^{q} \frac{Y_{.j.}^2}{n_{.j}} - \frac{Y_{...}^2}{N}.$$

4. Other models. We still assume that

$$Y_{ijk_{ij}} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk_{ij}}.$$

For the Type II model we assume that the τ_i's, β_j's $(\tau\beta)_{ij}$'s and $\epsilon_{ijk_{ij}}$'s are NID with zero means and variances σ_τ^2, σ_β^2, $\sigma_{\tau\beta}^2$, and σ^2, respectively. For the Type III model we assume that the τ_i's, β_j's, and $(\tau\beta)_{ij}$'s come from finite independent populations of size $P > p$, $Q > q$, and PQ, respectively, with zero means and variances

$$\sigma_\tau^2 = \frac{\sum_{i=1}^{P} \tau_i^2}{P-1}, \qquad \sigma_\beta^2 = \frac{\sum_{j=1}^{Q} \beta_j^2}{Q-1}, \qquad \sigma_{\tau\beta}^2 = \frac{\sum_{i,j}^{P,Q} (\tau\beta)_{ij}^2}{(P-1)(Q-1)}.$$

The assumption of zero means implies that

$$\sum_{i=1}^{P} \tau_i = 0, \qquad \sum_{j=1}^{Q} \beta_j = 0, \qquad \sum_{i,j}^{P,Q} (\tau\beta)_{ij} = 0,$$

and, in addition, we assume that

$$\sum_{i=1}^{P} (\tau\beta)_{ij} = \sum_{j=1}^{Q} (\tau\beta)_{ij} = 0.$$

For the mixed model we may assume that the τ_i's, β_j's, and $(\tau\beta)_{ij}$'s are of any of the types described above. In addition when the τ_i's, say, are of Type I and the β_i's of Type II, Anderson and Kempthorne ([1], [2]) have shown that it is desirable to assume that, corresponding to each β_j, there exists a population of $(\tau\beta)_{ij}$'s consisting of p elements such that

$$\sum_{i=1}^{p} (\tau\beta)_{ij} = 0, \qquad \sigma_{\tau\beta}^2 = \frac{\sum_{i=1}^{p} (\tau\beta)_{ij}^2}{p-1}.$$

If the τ_i's came from a Type III population, we would replace p by P in the above definitions, and if the roles of the τ_i's and β_j's were interchanged, we

would interchange i and j and replace p by q. We always assume the $\epsilon_{ijk_{ij}}$'s are NID $(0, \sigma^2)$.

5. The expected values of the sums of squares. In every case we shall arbitrarily begin with the sums of squares obtained for the Type I model. To determine their expected values, we shall make use of the following theorem which is a slight generalization of one stated by Tukey [11].

THEOREM B.

If $y_1, y_2, \cdots, y_p$ have means $\mu_1, \mu_2, \cdots, u_p$, variances $\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2$, and every pair has the same covariance, λ, then

$$E\left\{\sum_{i=1}^{p} n_i(y_i - \bar{y}_.)^2\right\} = \sum_{i=1}^{p} n_i(\mu_i - \bar{\mu}_.)^2 + \sum_{i=1}^{p} n_i\left(1 - \frac{n_i}{N}\right)(\sigma_i^2 - \lambda)$$

where

$$\bar{y}_. = \frac{\sum_{i=1}^{p} n_i y_i}{N}, \qquad \bar{\mu}_. = \frac{\sum_{i=1}^{p} n_i \mu_i}{N}, \qquad N = \sum_{i=1}^{p} n_i .$$

We find that

$$SST = \sum_{i=1}^{p} n_{i.}(w_i - \bar{w}_.)^2, \qquad SSB = \sum_{j=1}^{q} n_{.j}(y_j - \bar{y}_.)^2,$$

$$SS(TB) = \sum_{i,j}^{p,q} n_{ij}(z_i - \bar{z}_.)^2, \qquad SSE = \sum_{i,j,k_{ij}}^{p,q,n_{ij}} (\epsilon_{ijk_{ij}} - \bar{\epsilon}_{ij.})^2,$$

where

$$w_i = \tau_i + (\overline{\tau\beta})_{i.} + \bar{\epsilon}_{i..}, \qquad y_j = \beta_j + (\overline{\tau\beta})_{.j} + \bar{\epsilon}_{.j.},$$

$$z_i = (\tau\beta)_{ij} - (\overline{\tau\beta})_{i.} + \bar{\epsilon}_{ij.} - \bar{\epsilon}_{i..}$$

and

$$\bar{\tau}_. = \frac{\sum_{i=1}^{p} n_{i.}\tau_i}{N}, \qquad \bar{\beta}_. = \frac{\sum_{j=1}^{q} n_{.j}\beta_j}{N}, \qquad (\overline{\tau\beta})_{i.} = \frac{\sum_{j=1}^{q} n_{.j}(\tau\beta)_{ij}}{N},$$

$$(\overline{\tau\beta})_{.j} = \frac{\sum_{i=1}^{p} n_{i.}(\tau\beta)_{ij}}{N}, \qquad (\overline{\tau\beta})_{..} = \frac{\sum_{i,j}^{p,q} n_{ij}(\tau\beta)_{ij}}{N}.$$

In order to apply Theorem B, we need the variances and covariances of the w_i, y_j, and z_i in a form that does not depend on the form of the model. By using the methods employed by Bennett and Franklin [3], we find for the Type III model that

$$\mu_i = E(w_i) = (1 - \delta_\tau)\tau_i, \qquad \bar{\mu}_. = 0,$$

$$\sigma_i^2 = \delta_\tau\left(1 - \frac{1}{P}\right)\sigma_\tau^2 + \delta_{\tau\beta}\left(1 - \frac{1}{P}\right)\frac{1}{N^2}\left(\sum_{j=1}^{q} n_{.j}^2 - \frac{N^2}{Q}\right)\sigma_{\tau\beta}^2 + \frac{\sigma^2}{n_{i.}}.$$

$$\lambda = -\delta_\tau \frac{\sigma_\tau^2}{P} - \delta_{\tau\beta} \frac{1}{PN^2} \left(\sum_{j=1}^{q} n_{\cdot j}^2 - \frac{N^2}{Q} \right) \sigma_{\tau\beta}^2 ,$$

where $\delta_\tau = 0$ if the τ_i's come from a Type I population, $\delta_\tau = 1$ otherwise, and a similar definition holds for $\delta_{\tau\beta}$.

Application of Theorem B enables us to find $E(SST)$ and division by $p - 1$ gives us

$E(MST)$

$$= \sigma^2 + \frac{\delta_{\tau\beta} a}{(p-1)N^2} \left(\sum_{j=1}^{q} n_{\cdot j}^2 - \frac{N^2}{Q} \right) \sigma_{\tau\beta}^2 + \frac{\delta_\tau a}{p-1} \sigma_\tau^2 + \frac{(1-\delta_\tau)}{p-1} \sum_{i=1}^{p} n_{i\cdot} \tau_i^2 ,$$

where

$$a = N - \frac{1}{N} \sum_{i=1}^{p} n_{i\cdot}^2 .$$

Similarly

$E(MSB)$

$$= \sigma^2 + \frac{\delta_{\tau\beta} b}{(q-1)N^2} \left(\sum_{i=1}^{p} n_{i\cdot}^2 - \frac{N^2}{P} \right) \sigma_{\tau\beta}^2 + \frac{\delta_\beta b}{q-1} \sigma_\beta^2 + \frac{(1-\delta_\beta)}{q-1} \sum_{i=1}^{q} n_{\cdot\beta} \beta_j^2 ,$$

where

$$b = N - \frac{1}{N} \sum_{j=1}^{q} n_{\cdot j}^2 ,$$

and

$$E[MS(TB)] = \sigma^2 + \frac{\delta_{\tau\beta} ab}{(p-1)(q-1)N} \sigma_{\tau\beta}^2 + \frac{(1-\delta_{\tau\beta}) \sum_{i,j}^{p,q} n_{ij} (\tau\beta)_{ij}^2}{(p-1)(q-1)}$$

Finally, by the theory for the Type I model, we know that SSE is distributed as $\chi^2\sigma^2$ with $N - pq$ degrees of freedom and hence $E(MSE) = \sigma^2$. We also note that, if all the $n_{ij} = 1$, $SSE = 0$ and it is impossible to carry out any of the F tests which involve division by this quantity.

6. Models with no interaction. In this case, for the Type I model,

$$Y_{ijk_{ij}} = \mu + \tau_i + \beta_j + \epsilon_{ijk_{ij}} .$$

We find, as in Sec. 2, that

$$m = \bar{Y}_{\cdots} , \qquad t_i = \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots} , \qquad b_j = \bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdots} ,$$

and

$$SSE_1 = \sum_{i,j,k_{ij}}^{p,q,n_{ij}} (Y_{ijk_{ij}} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdot\cdot})^2$$

of that section plays the role of SSE. We also saw in Sec. 2 that

$$SSE_1 = SSE + SS(TB)$$

so, if we had accepted

$$H_1: (\tau\beta)_{ij} = 0,$$

and decided to change models in midstream, all that would be necessary to obtain SSE_1 would be to pool the interaction and error sums of squares. We find the same expressions for SST and SSB as in Sec. 3 and that the degrees of freedom associated with SSE_1 are those obtained by pooling the degrees of freedom associated with $SS(TB)$ and SSE. To obtain the expected values of MST and MSB one need only omit the terms involving the $(\tau\beta)_{ij}$, and it is easily verified that $E(MSE_1) = \sigma^2$. A discussion as to when pooling is desirable is to be found in Bechhofer's thesis [2] and in a paper by Bozivich, Bancroft and Hartley [4].

7. Distributions of the sums of squares. Corresponding to the hypotheses

$$H_1: (\tau\beta)_{ij} = 0, \qquad H_2: \tau_i = 0, \qquad H_3: \beta_j = 0,$$

we have the hypotheses

$$\sigma_{\tau\beta} = 0, \qquad \sigma_\tau = 0, \qquad \sigma_\beta = 0,$$

if the corresponding variables are from other than a Type I population. Then since the populations have zero means, it follows that the corresponding variables are equal to zero. We have already referred to the tests for the above hypotheses for the Type I model at the end of Sec. 2. If there is no interaction term, subject to H_2 and H_3, the sums of squares SST and SSB reduce to the corresponding expressions for the Type I model and the tests of Sec. 6 apply no matter which model we may be considering. If there is an interaction term, the same argument shows that the Type I test can be used for H_1. Thus our problem is reduced to testing H_2 and H_3 when there is interaction and we are not dealing with a Type I model.

We first consider the Type II model where the parameters are NID with zero means and variances σ_τ^2, σ_β^2, and $\sigma_{\tau\beta}^2$. If all the n_{ij}'s are equal to n, say, using methods similar to those of Mood [6], it can be shown that SST, SSB, $SS(TB)$ and SSE are independently distributed as $\chi^2 E(MST)$, $\chi^2 E(MSB)$, $\chi^2 E[MS(TB)]$, and $\chi^2\sigma^2$ with $p - 1$, $q - 1$, $(p - 1)(q - 1)$, and $N - pq$ degrees of freedom, respectively. These results hold independent of the validity of H_1, H_2, and H_3. Since some of the details differ from those given by Mood we shall outline the proof of the above results.

The theory for the Type I model shows that

$$t_i = \bar{Y}_{i..} - \bar{Y}_{...}, \qquad bj = \bar{Y}_{.j.} - \bar{Y}_{...}, \qquad (tb)_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...},$$

$i = 1, 2, \cdots, p - 1; j = 1, 2, \cdots, q - 1$, are distributed independently of

$$SSE = \sum_{i,j,k_{ij}}^{p,q,n_{ij}} (\epsilon_{ijk_{ij}} - \bar{\epsilon}_{ij.})^2.$$

Therefore any function of these statistics is distributed independently of SSE, and, in particular, this holds for t_p, b_q, $(tb)_{pj}$ and $(tb)_{iq}$. These results hold for the particular case where $Y_{ijk} = \epsilon_{ijk}$. Hence

$$\bar{\epsilon}_{i..} - \bar{\epsilon}_{...}, \qquad \bar{\epsilon}_{.j.} - \bar{\epsilon}_{...}, \qquad \bar{\epsilon}_{ij.} - \bar{\epsilon}_{i..} - \bar{\epsilon}_{.j.} + \bar{\epsilon}_{...},$$

$i = 1, 2, \cdots, p; j = 1, 2, \cdots, q$, are distributed independently of SSE. It may be shown that any variable of the above three types is independent of any variable of the other two types by computing the appropriate covariances. We know that

$$SST = qn \sum_{i=1}^{p} (w_i - \bar{w}_.)^2$$

where

$$w_i = \tau_i + (\overline{\tau\beta})_{i.} + \bar{\epsilon}_{i..}, \qquad E(w_i) = 0, \qquad \text{var}\,(w_i) = \sigma_\tau^2 + \frac{\sigma_{\tau\beta}^2}{q} + \frac{\sigma^2}{qn},$$

$$\text{cov}\,(w_{i'}, w_i) = 0, \qquad E(MST) = \sigma^2 + n\sigma_{\tau\beta}^2 + qn\sigma_\tau^2.$$

It follows that $SST / E(MST)$ has a χ^2 distribution with $p - 1$ degrees of freedom. Similarly SSB is distributed as $\chi^2 E(MSB)$ with $q - 1$ degrees of freedom.

Consider the three sets of variables

$$(\overline{\tau\beta})_{i.} - (\overline{\tau\beta})_{..}, \qquad (\overline{\tau\beta})_{.j} - (\overline{\tau\beta})_{..}, \qquad (\overline{\tau\beta})_{ij} - (\overline{\tau\beta})_{i.} - (\overline{\tau\beta})_{.j} + (\overline{\tau\beta})_{..}.$$

As with the ϵ_{ijk}'s, it may be shown that any variable of the above three types is independent of any variable of the other two types. Then it follows that the three sets of variables

$$w_i - \bar{w}_., \qquad y_j - \bar{y}_., \qquad z_i - \bar{z}_.$$

are independently distributed and hence so are SST, SSB, and $SS(TB)$.

If, in the results for the Type I model, we set μ, the τ_i's and the β_j's equal to zero and assume the $(\tau\beta)_{ij}$'s are NID $(0, \sigma_{\tau\beta}^2)$,

$$Y_{ijk} = (\tau\beta)_{ij} + \epsilon_{ijk}, \qquad \bar{Y}_{ij.} = (\tau\beta)_{ij} + \bar{\epsilon}_{ij.},$$

$$E(\bar{Y}_{ij.}) = 0, \qquad \text{var}\,(\bar{Y}_{ij.}) = \sigma_{\tau\beta}^2 + \frac{\sigma^2}{n},$$

and the $\bar{Y}_{ij.}$'s are independent. We carry out an analysis of variance on the $\bar{Y}_{ij.}$'s according to the model of Sec. 6, where there is no interaction, with the N of *that* section equal to pq and $n_{ij} = 1$, to obtain

$$SSE_1 = SSE + SS(TB) = \sum_{i,j}^{p,q} (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2$$

since, under these conditions, the Y_{ijk} of that section is equal to $\bar{Y}_{ij\cdot}$. The theory of Sec. 6, when we replace σ^2 by $\sigma_{\tau\beta}^2 + \sigma^2/n$, tells us that SSE_1 is distributed as $\chi^2(n\sigma_{\tau\beta}^2 + \sigma^2)$ and hence

$$SS(TB) = \sum_{i,j}^{p,q} n(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot} + \bar{Y}_{\cdot\cdot\cdot})^2$$

is distributed as $\chi^2(n\sigma_{\tau\beta}^2 + \sigma^2)$ with $pq - p - q + 1 = (p-1)(q-1)$ degrees of freedom. It then follows that the appropriate tests for H_1, H_2, H_3 are given by the statistics

$$F_1 = \frac{MS(TB)}{MSE}, \qquad F_2 = \frac{MST}{MS(TB)}, \qquad F_3 = \frac{MSB}{MS(TB)},$$

respectively. A proof of these results is also outlined by Anderson and Bancroft [1].

The above results are for the case where $n_{ij} = n$. If this condition does not hold, we can no longer say that F_2 and F_3 have the F distribution. This may be shown by considering the special case where $p = 3$, $q = 2$, $n_{11} = n_{12} = 1$, $n_{21} = n_{22} = 2$, $n_{31} = n_{32} = 3$ and $N = 12$. Then the moment generating function of SST is

$$[1 - 4(11x + 3\sigma^2)t/3 + 4(36x^2 + 22x\sigma^2 + 3\sigma^4)t^2/3]^{-1/2}$$

where $x = \sigma_\tau^2 + \sigma_{\tau\beta}^2/2$. This is not the moment generating function of a variable of the form $c\chi^2$ unless $x = 0$. Thus there is no hope of F_2 having an F distribution and a similar argument holds for F_3.

For a Type III model with interaction we cannot expect to obtain the distributions necessary for F tests of H_2 and H_3 since the $(\tau\beta)_{ij}$'s are not normally distributed. An approach similar to the one given above could be used in the case of the mixed model.

8. The two-way nested classifications. This model is discussed by Bennett and Franklin [3]. We assume that

$$Y_{i\beta_{(i)}j} = \mu + \tau_i + \beta_{j(i)} + \epsilon_{i\beta_{(i)}j},$$

$$\sum_{i=1}^{p} n_{i\cdot}\tau_i = 0, \qquad \sum_{j=1}^{q} n_{\cdot j}\beta_{j(i)} = 0, \qquad i = 1, 2, \cdots, p.$$

We test two hypotheses,

$$H_1: \beta_{j(i)} = 0, \qquad i = 1, 2, \cdots, p; j = 1, 2, \cdots, q,$$

and

$$H_2: \tau_i = 0, \qquad i = 1, 2, \cdots, p,$$

using the statistics, for the Type I model,

$$F_1 = \frac{MSB}{MSE} \text{ (with } p(q-1) \text{ and } N - pq \text{ degrees of freedom)},$$

and

$$F_2 = \frac{MST}{MSE} \text{ (with } p - 1 \text{ and } N - pq \text{ degrees of freedom),}$$

where SST and SSE have the values given earlier and

$$SSB = \sum_{i,j}^{p,q} n_{ij}(\bar{Y}_{ij.} - \bar{Y}_{i..})^2 = \sum_{i,j}^{p,q} \frac{Y_{ij.}^2}{n_{ij}} - \sum_{i=1}^{p} \frac{Y_{i..}^2}{n_{i.}}.$$

The Type II model is defined as before but, in the case of the Type III model, we assume that the τ_i's come from a finite population of size P, mean zero, and variance

$$\sigma_\tau^2 = \frac{\sum_{i=1}^{P} \tau_i^2}{P - 1}$$

while the $\beta_{j(i)}$'s come from P populations of size Q, corresponding to the different values of i, these populations being independent of each other and the population of τ_i's, with zero means and common variance

$$\sigma_\beta^2 = \frac{\sum_{j=1}^{Q} \beta_{j(i)}^2}{Q - 1}.$$

The expected values of the mean squares are

$$E(MST) = \sigma^2 + \frac{\delta_\beta a}{(p-1)N^2}\left[\sum_{j=1}^{q} n_{.j}^2 - \frac{N^2}{Q}\right]\sigma_\beta^2 + \frac{\delta_\tau a}{p-1}\sigma_\tau^2 + \frac{(1-\delta_\tau)}{p-1}\sum_{i=1}^{p} n_{i.}\tau_i^2,$$

$$E(MSB) = \sigma^2 + \frac{\delta_\beta b}{p(q-1)}\sigma_\beta^2 + \frac{(1-\delta_\beta)}{p(q-1)}\sum_{i,j}^{p,q} n_{ij}\beta_{j(i)}^2, \qquad E(MSE) = \sigma^2,$$

where the δ's have the meaning assigned in Sec. 5 and

$$a = N - \frac{\sum_{i=1}^{p} n_{i.}^2}{N}, \qquad b = N - \frac{\sum_{j=1}^{q} n_{.j}^2}{N}$$

Examination of MSB indicates that, no matter what model we may use, we may test the hypothesis H_1 by the statistic F_1 given earlier in this section. For a Type II model with $n_{ij} = n$, we test H_2 with the statistic

$$F = \frac{MST}{MSB}.$$

For other cases an approximate method must be used such as is given elsewhere in the literature [7].

REFERENCES

[1] R. L. Anderson and T. A. Bancroft, *Statistical Theory in Research*, McGraw-Hill, New York, 1952, pp. 177–182, 234, 314–319, and 340.

[2] R. E. Bechhofer, "The effect of preliminary tests of significance on the size and power of certain tests of univariate linear hypotheses," Ph.D. Thesis, Columbia University Library.

[3] C. A. Bennett and N. L. Franklin, *Statistical Analysis in Chemistry and the Chemical Industry*, John Wiley & Sons, New York, 1954, pp. 470–477.

[4] Helen Bozivich, T. A. Bancroft and H. O. Hartley, "Power of analysis of variance test procedures for certain incompletely specified models, I," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 1017–1043.

[5] H. B. Mann, *Analysis and Design of Experiments*, Dover, New York, 1949, pp. 39–40.

[6] A. M. Mood, *Introduction to the Theory of Statistics*, McGraw-Hill, New York, 1950, p. 344.

[7] F. E. Satterthwaite, "An approximate distribution of estimates of variance components," *Biometrics*, Vol. 2 (1946), pp. 110–114.

[8] H. F. Smith, "Analysis of variance with unequal but proportionate numbers of observations in the sub-classes of a two-way classification," *Biometrics*, Vol. 7 (1951), pp. 70–74.

[9] G. W. Snedecor and G. M. Cox, "Disproportionate Subclass Numbers in Tables of Multiple Classification," Iowa State College, Agr. Expt. Bull. 180 (1935).

[10] G. W. Snedecor, *Statistical Methods*, 4th ed., Iowa State College Press, Ames, Iowa, pp. 281–284.

[11] J. W. Tukey, "Dyadic anova, an analysis of variance for vectors," *Human Biology*, Vol. 21 (1949), pp. 65–110.

[12] O. Kempthorne, *The Design and Analysis of Experiments*, John Wiley & Sons, New York, 1952, p. 574.

SUMS OF INDEPENDENT TRUNCATED RANDOM VARIABLES[1]

By J. M. Shapiro

The Ohio State University

1. Summary and introduction. Let (x_{nk}), $(k = 1, 2, \cdots, k_n ; n = 1, 2, \cdots)$ be a double sequence of infinitesimal (i.e. $\lim_{n\to\infty} \max_{1\leqq k\leqq k_n} P\{|x_{nk}| > \epsilon\} = 0$ for every $\epsilon > 0$) random variables such that for each n, $x_{n1}, \cdots, x_{nk_n}$ are independent. Let $S_n = x_{n1} + \cdots + x_{nk_n}$ and let $F_n(x)$ be the distribution function of S_n. For any $a > 0$ let the random variables x_{nk}^a be defined by

$$x_{nk}^a = \begin{cases} x_{nk}, & \text{if } -a < x_{nk} \leqq a, \\ 0, & \text{otherwise,} \end{cases}$$

and let $F_n^a(x)$ be the distribution function of $S_n^a = x_{n1}^a + \cdots + x_{nk_n}^a$. In the next section certain necessary and sufficient conditions are given for $F_n^a(x)$ to converge $(n \to \infty)$ to a limiting distribution and in particular it is shown that if $F_n^a(x)$ converges to $F(x)$, then $F(x)$ has finite moments of all orders. In Sec. 3 it is shown that if $F_n^a(x)$ converges to $F(x)$, then for each positive integer k the kth moment of $F_n^a(x)$ approaches the kth moment of $F(x)$ as $n \to \infty$.

We shall call the random variables (x_{nk}) a truncated system if there exists a $b > 0$ independent of k and n such that $P\{|x_{nk}| > b\} = 0$. We note that if we start with a truncated system we can choose $a > 0$ such that $x_{nk}^a = x_{nk}$.

2. Conditions for convergence. Since the random variables (x_{nk}) are infinitesimal and independent within each row, it is clear that the random variables (x_{nk}^a) are also. From a well-known theorem of Khintchine, (c.f. [1]), it follows that for the weak convergence of $F_n(x)$ (or $F_n^a(x)$) to a limiting distribution $F(x)$, $F(x)$ must be infinitely divisible.

Let $F(x)$ be any infinitely divisible distribution function and let $\varphi(t)$ be its characteristic function. According to the formulas of Levy and Khintchine [1] for the representation of the characteristic function of an infinitely divisible distribution we have

$$\begin{aligned} \log \varphi(t) &= i\gamma t + \int_{-\infty}^{\infty} \left(e^{iut} - 1 - \frac{iut}{1+u^2} \right) \frac{1+u^2}{u^2}\, dG(u) \\ &= i\gamma(\tau)t - b^2t^2/2 + \int_{-\infty}^{-\tau} (e^{iut} - 1)\, dM(u) \\ &\quad + \int_{\tau}^{\infty} (e^{iut} - 1)\, dN(u) + \int_{-\tau}^{0^-} (e^{iut} - 1 - iut)\, dM(u) \\ &\quad + \int_{0^+}^{\tau} (e^{iut} - 1 - iut)\, dN(u), \end{aligned} \tag{2.1}$$

Received September 28, 1956.

[1] Presented to the American Mathematical Society on February 26, 1955 and December 29, 1955.

where $G(u)$ is bounded nondecreasing function ($G(-\infty) = 0$), γ a real constant,

$$M(u) = \int_{-\infty}^{u} \frac{1+z^2}{z^2}\, dG(z) \quad \text{for} \quad u < 0,$$

$$N(u) = -\int_{u}^{\infty} \frac{1+z^2}{z^2}\, dG(z) \quad \text{for} \quad u > 0, \tag{2.2}$$

$$b^2 = G(+0) - G(-0) \quad \text{and} \quad \gamma(\tau) = \gamma + \int_{|u|<\tau} u\, dG(u) - \int_{|u|\geq\tau} \frac{1}{u}\, dG(u),$$

and where τ and $-\tau$ are continuity points of $N(u)$ and $M(u)$ respectively.

Let $F_{nk}(x)$ and $F^a_{nk}(x)$ be the distribution functions of x_{nk} and x^a_{nk} respectively. From the definition of x^a_{nk} we note

$$F^a_{nk}(x) = \begin{cases} 0, & \text{for } x \leqq -a, \\ F_{nk}(x) - F_{nk}(-a), & \text{for } -a \leqq x < 0, \\ F_{nk}(x) + 1 - F_{nk}(a), & \text{for } 0 \leqq x \leqq a, \\ 1, & \text{for } x \geqq a. \end{cases} \tag{2.3}$$

The following theorem (c.f. [1], p. 124) will be needed.

THEOREM 1. *In order that the distribution functions of the sums* $S_1 = x_{n1} + \cdots + x_{nk_n}$ *of independent infinitesimal, random variables converge to the distribution function* $F(x)$, *it is necessary and sufficient that*:

(1) *At continuity points of* $M(u)$ *and* $N(u)$

$$\lim_{n\to\infty} \sum_{k=1}^{k_n} F_{nk}(x) = M(x), \quad \text{for} \quad x < 0,$$

$$\lim_{n\to\infty} \sum_{k=1}^{k_n} (F_{nk}(x) - 1) = N(x), \quad \text{for} \quad x > 0;$$

(2) $$\lim_{\epsilon\to 0} \overline{\lim_{n\to\infty}} \sum_{k=1}^{k_n} \left\{ \int_{|x|<\epsilon} x^2\, dF_{nk}(x) - \left(\int_{|x|<\epsilon} x\, dF_{nk}(x) \right)^2 \right\} =$$

$$\lim_{\epsilon\to 0} \underline{\lim_{n\to\infty}} \sum_{k=1}^{k_n} \left\{ \int_{|x|<\epsilon} x^2\, dF_{nk}(x) - \left(\int_{|x|<\epsilon} x\, dF_{nk}(x) \right)^2 \right\} = b^2;$$

(3) $$\lim_{n\to\infty} \sum_{k=1}^{k_n} \int_{|x|<\tau} x\, dF_{nk}(x) = \gamma(\tau),$$

where $M(u)$, $N(u)$, b^2, *and* $\gamma(\tau)$ *are given by* (2.1) *and* (2.2).

Now using the notation of (2.1) we have the following theorem.

THEOREM 2. *If for some* $a > 0$ $F^a_n(x)$ *converges to* $F(x)$, *then the function* $G(u)$ *is nonincreasing for* $u > a$ *and for* $u < -a$.

Proof. Since $F^a_n(x)$ converges to $F(x)$, according to Theorem 1 we know that at continuity points of $M(u)$ and $N(u)$

$$\lim_{n\to\infty} \sum_{k=1}^{k_n} F_{nk}^a(x) = M(x) \quad \text{for} \quad x < 0 \quad \text{and}$$

(2.4)

$$\lim_{n\to\infty} \sum_{k=1}^{k_n} (F_{nk}^a(x) - 1) = N(x) \quad \text{for} \quad x > 0.$$

Thus from (2.3) and (2.4) since $M(u)$ and $N(u)$ are nondecreasing functions, we see that $M(u) = 0$ for $u < -a$ and $N(u) = 0$ for $u > a$. Using (2.2) the conclusion of the theorem follows.

Now given $F(x)$ infinitely divisible define (using the notation of (2.1) and (2.2)) for any $a > 0$, $\pm a$ continuity points of $G(u)$,

$$G^a(u) = \begin{cases} 0, & \text{for } u \leqq -a, \\ G(u) - G(-a), & \text{for } -a \leqq u \leqq a, \\ G(a) - G(-a), & \text{for } u \geqq a, \end{cases} \tag{2.5}$$

$$\gamma^a = \gamma - \int_{|u|>a} \frac{1}{u}\, dG(u),$$

and let $F^a(x)$ be the (infinitely divisible) distribution given by (2.1) using the function $G^a(u)$ and the constant γ^a. We note that $F^a(x)$ is also given by (2.1) using the function $M^a(u)$ and $N^a(u)$ defined by

$$M^a(u) = \begin{cases} 0, & \text{for } -\infty < u < -a, \\ M(u) - M(-a), & \text{for } -a \leqq u < 0, \end{cases}$$

(2.6)

$$N^a(u) = \begin{cases} 0, & \text{for } a < u < \infty, \\ N(u) - N(a), & \text{for } 0 < u \leqq a, \end{cases}$$

(with b^2 unchanged) and

$$\gamma^a(a) = \begin{cases} \gamma(\tau) & \text{for} \quad \tau \leqq a, \\ \gamma(\partial) & \text{for} \quad \tau > a. \end{cases}$$

(With this notation we have the following theorem.

THEOREM 3. *If $F_n(x)$ converges to $F(x)$, then for any $a > 0$ ($\pm a$ continuity points of $G(u)$) $F_n^a(x)$ converges to $F^a(x)$. In particular, if $G(u)$ is nonincreasing outside of the interval $[-a, a]$ then $F_n^a(x)$ converges to $F(x)$.*

Proof. Since $F_n(x)$ converges to $F(x)$, parts (1), (2), and (3) of Theorem 1 hold. We note that continuity points of $M(u)$ and $N(u)$ coincide with those of $G(u)$ so that $-a$ and a are continuity points of $M(u)$ and $N(u)$ respectively. From (2.3) and (2.6) it follows that at continuity points of $M^a(u)$ and $N^a(u)$

$$\lim_{n\to\infty} \sum_{k=1}^{k_n} F_{nk}^a(x) = M^a(x) \quad \text{and} \quad \lim_{n\to\infty} \sum_{k=1}^{k_n} (F_{nk}^a(x) - 1) = N^a(x),$$

for $x < 0$ and $x > 0$ respectively. Also it is clear that part (2) of Theorem 1

holds with $F_{nk}(x)$ replaced by $F^a_{nk}(x)$ and that

$$\lim_{n\to\infty} \sum_{k=1}^{k_n} \int_{|x|<\tau} x\, dF^a_{nk}(x) = \gamma^a(\tau).$$

Thus from the sufficiency of Theorem 1 we see that $\lim_{n\to\infty} F^a_n(x) = F^a(x)$ (at continuity points of $F^a(x)$). We note that if $G(u)$ is nonincreasing outside of $[-a, a]$ then $F^a(x) = F(x)$. This proves Theorem 3.

Combining Theorems 2 and 3 we can state the following theorem.

THEOREM 4. *If $F_n(x)$ converges to $F(x)$ and if $\pm a$ are continuity points of $G(u)$, then a necessary and sufficient condition for $F^a_n(x)$ to converge to $F(x)$ is that $G(u) = G(+\infty)$ for $u \geqq a$ and $G(u) = G(-\infty) = 0$ for $u \leqq -a$.*

THEOREM 5. *If $F^a_n(x)$ converges to $F(x)$, then $F(x)$ has finite moments of all orders.*

Proof. By Theorem 2 we know that $G(u)$ is nonincreasing outside of the interval $[-a, a]$. In particular it follows that $\int_{-\infty}^{\infty} x^n\, dG(x) < \infty$ for all n. By the result of [2] it follows that $F(x)$ has finite moments of all orders.

We remark that if the system (x_{nk}) is a truncated system we have the following analogues of Theorems 2 and 5.

THEOREM 2a. *If $F_n(x)$ converges to $F(x)$, then the function $G(u)$ is nonincreasing for $u > a$ and for $u < -a$.*

THEOREM 5a. *If $F_n(x)$ converges to $F(x)$, then $F(x)$ has finite moments of all orders.*

3. Convergence of moments. In the remainder of this paper we shall assume that (x_{nk}) is a truncated system. If this is not the case, the following results apply to the system (x^a_{nk}) previously discussed.

In view of Theorem 5a it is natural to consider the question of the convergence of moments of the distribution function $F_n(x)$ of the random variable S_n to the moments of $F(x)$. The principle result of this section is contained in the following theorem.

THEOREM 6. *If (x_{nk}) is a truncated system, and if $F_n(x)$ converges to $F(x)$, then*

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} x^k\, dF_n(x) = \int_{-\infty}^{\infty} x^k\, dF(x),$$

for every positive integer k.

The author first proved this theorem in the special case where $F(x)$ was the Poisson distribution (see *Bull. Amer. Math. Soc.*, Vol. 61, Abstract No. 435) and where $k = 2$. This more general form was obtained at a later date (*Bull. Amer. Math. Soc.*, Vol. 62, Abstract No. 264).

The proof of Theorem 6 requires several lemmas which we state and prove below.

Using the same notation as in section 2, according to the result of [2] we know that

$$\int_{-\infty}^{\infty} x^{2k}\, dF(x) < \infty \Leftrightarrow \int_{-\infty}^{\infty} x^{2k}\, dG(x) < \infty,$$

and assuming $F(x)$ has finite moments of all orders that,

$$\chi_1 = \gamma + \int_{-\infty}^{\infty} u\, dG(u) \quad \text{and} \quad \chi_r = \int_{-\infty}^{\infty} (u^{r-2} + u^r)\, dG(u), \tag{3.1}$$

where χ_r is the rth semi-invariant of $F(x)$. In particular letting μ be the mean and σ^2 the variance of $F(x)$, we see

$$\mu = \gamma + \int_{-\infty}^{\infty} u\, dG(u) \quad \text{and} \quad \sigma^2 = G(+\infty) + \int_{-\infty}^{\infty} u^2\, dG(u). \tag{3.2}$$

LEMMA 1. *Under the hypothesis of Theorem* 6, $\lim_{n\to\infty} \sigma^2(S_n) = \sigma^2$, *where* $\sigma^2(S_n)$ *is the variance of* S_n *and* σ^2 *is the variance of* $F(x)$.

Proof. Since $F_n(x)$ converges to $F(x)$ by Theorem 1, page 112 of [1], we have $G_n(x) \equiv \sum_{k=1}^{k_n} \int_{-\infty}^{x} u^2/(1+u^2)\, dF_{nk}(u + \alpha_1) \to G(x)$ as $n \to \infty$ at all continuity points of $G(u)$ and also $G_n(+\infty) \to G(+\infty)$, where $\alpha_{nk} = \int_{|x|<\tau} x\, dF_{nk}(x)$, ($\tau > 0$ an arbitrary positive constant). (Remark. By hypothesis $P\{|x_{nk}| > a\} = 0$ for some $a > 0$. We may and do take $\tau > a$ so that $\alpha_{nk} = \mu_{nk}$ = mean of x_{nk}. Hence in the remainder of the proof we assume $\alpha_{nk} = \mu_{nk}$.) Now since $x^3/(1+x^2) = x^3 - [x^4/(1+x^2)]$, we see

$$\begin{aligned} G_n(+\infty) &= \sum_{k=1}^{k_n} \int_{-\infty}^{\infty} x^3\, dF_{nk}(x + \mu_{nk}) \\ &\quad - \sum_{k=1}^{k_n} \int_{-\infty}^{\infty} x^4/(1+x^2)\, dF_{nk}(x + \mu_{nk}) \to G(+\infty) \quad \text{as} \quad n \to \infty. \end{aligned} \tag{3.3}$$

Also,

$$\begin{aligned} \int_{-\infty}^{\infty} x^2\, dG_n(x) &= \int_{-\infty}^{\infty} x^2\, d \sum_{k=1}^{k_n} \int_{-\infty}^{x} \frac{u^2}{1+u^2}\, dF_{nk}(u + \mu_{nk}) \\ &= \sum_{k=1}^{k_n} \int_{-\infty}^{\infty} x^4/(1+x^2)\, dF_{nk}(u + \mu_{nk}). \end{aligned} \tag{3.4}$$

By Theorem 2a, $G(x)$ is nonincreasing outside of some interval. Now since the random variables are infinitesimal it follows that $\lim_{n\to\infty} \max_{1\le k\le k_n} |\alpha_{nk}| = 0$. Thus since $P\{|x_{nk}| > a\} = 0$ for some $a > 0$, we know that there exists an $A > 0$ such that $G(x)$ and $G_n(x)$ are nonincreasing for $x < -A$ and $x > A$ ($n = 1, 2, \cdots$). Therefore by Helly's convergence theorem

$$\int_{-\infty}^{\infty} x^k\, dG_n(x) = \int_{-A}^{A} x^k\, dG_n(x) \to \int_{-A}^{A} x^k\, dG(x) = \int_{-\infty}^{\infty} x^k\, dG(x) \tag{3.5}$$

as $n \to \infty$. Letting $k = 2$ and using (3.3) and (3.4) we see

$$\lim_{n\to\infty} \sum_{k=1}^{k_n} \int_{-\infty}^{\infty} x^2\, dF_{nk}(x + \alpha_{nk}) = G(+\infty) + \int_{-\infty}^{\infty} x^2\, dG(x).$$

Now $x_{n1}, x_{n2}, \cdots, x_{nk_n}$ are for each n independent random variables and since $\alpha_{nk} = \mu_{nk}$, by virtue of (3.2) we see $\lim_{n\to\infty} \sigma^2(S_n) = \sigma^2$. This proves Lemma 1.

Having obtained this result we can now prove that the means μ_n of $F_n(x)$ approach the mean μ of $F(x)$.

LEMMA 2. *Under the hypothesis of Theorem* 6,

$$\mu_n = \int_{-\infty}^{\infty} x\, dF_n(x) \to \int_{-\infty}^{\infty} x\, dF(x) = \mu \quad \text{as} \quad n \to \infty$$

(*i.e., Theorem* 6 *holds for* $k = 1$.)

Proof. For the proof of this lemma we appeal to Theorem 2 of [1], page 100. Since the random variables (x_{nk}) are infinitesimal and since $\max_{1\leq k\leq k_n} |\mu_{nk}| = \max_{1\leq k\leq k_n} |\alpha_{nk}| \to 0$ as $n \to \infty$ we see that the random variables $(x_{nk} - \mu_{nk})$ are also infinitesimal. This together with Lemma 1 shows that the hypothesis of Theorem 2, page 100 of [1] is satisfied and hence we may conclude in particular that

$$\mu_n = \sum_{k=1}^{k_n} \int_{-\infty}^{\infty} x\, dF_{nk}(x) \to \gamma' \quad \text{as} \quad n \to \infty,$$

where γ' is the constant associated with Kolmogorov's formula for the characteristic function of the infinitely divisible distribution $F(x)$. But the constant of Kolmogorov's formula is the mean of the distribution (i.e. $\gamma' = \mu$). This proves Lemma 2.

LEMMA 3. *Under the hypothesis of Theorem* 6

$$\sum_{k=1}^{k_n} \int_{-\infty}^{\infty} x^r\, dF_{nk}(x + \mu_{nk}) - \chi_{r(n)} \to 0 \quad \text{as} \quad n \to \infty,$$

$r = 2, 3, \cdots$, *where* $\chi_{r(n)} = r$*th semi-invariant of* S_n.

Proof. We note that

$$\sum_{k=1}^{k_n} \int_{-\infty}^{\infty} x^r\, dF_{nk}(x + \mu_{nk}) - \chi_{r(n)} = 0 \quad \text{for} \quad r = 2, 3$$

and

$$\sum_{k=1}^{k_n} \int_{-\infty}^{\infty} x\, dF_{nk}(x) - \chi_{r(1)} = 0.$$

Let

$$\mu_n^{(r)} = \int_{-\infty}^{\infty} (x - \mu_n)^r\, dF_n(x)$$

and let $\mu_{nk}^{(r)} = \int_{-\infty}^{\infty} x^r\, dF_{nk}(x + \mu_{nk})$. Now since (x_{nk}) is a truncated system, and since $\max_{1\leq k\leq k_n} |\mu_{nk}| \to 0$ as $n \to \infty$ we see

$$\max_{1\leq k\leq k_n} \left| \int_{-\infty}^{\infty} x^r\, dF_{nk}(x + \mu_{nk}) \right| = \max_{1\leq k\leq k_n} \left| \int_{-A}^{A} x^r\, dF_{nk}(x + \mu_{nk}) \right|$$

for some $A > 0$. Now given $0 < \epsilon < 1$ we see

$$\max_k \left| \int_{-A}^{A} x^r\, dF_{nk}(x + \mu_{nk}) \right| \leqq \max_k \int_{\epsilon}^{\epsilon} |x|^r\, dF_{nk}(x + \mu_{nk})$$

$$+ \max_k \int_{\epsilon \leq |x| \leq A} |x|^r \, dF_{nk}(x + \mu_{nk}) \leqq \epsilon^r + A^r \max_k P\{|x - \mu_{nk}| \geqq \epsilon\}$$

and since $(x_{nk} - \mu_{nk})$ are infinitesimal we see

$$\lim_{n \to \infty} \max_{1 \leq k \leq k_n} \left| \int_{-\infty}^{\infty} x^r \, dF_{nk}(x + \mu_{nk}) \right| = 0, r = 2, 3, \cdots . \tag{3.6}$$

Also we see (for $r \geqq 2$)

$$\sum_{k=1}^{k_n} |\mu_{nk}^{(r)}| = \sum_{k=1}^{k_n} \left| \int_{-\infty}^{\infty} x^r \, dF_{nk}(x + \mu_{nk}) \right|$$

$$\leqq \sum_{k=1}^{k_n} \int_{-A}^{A} |x^r| \, dF_{nk}(x + \mu_{nk}) \leqq A^{r-2} \sum_{k=1}^{k_n} \int_{-A}^{A} x^2 \, dF_{nk}(x + \mu_{nk})$$

$$= A^{r-2}\sigma^2(S_n) \to A^{r-2}\sigma^2 \text{ as } n \to \infty$$

by Lemma 1. Hence

$$\sum_{k=1}^{k_n} |\mu_{nk}^{(r)}| \tag{3.7}$$

is bounded in n for $r = 2, 3 \cdots$. Let $\chi_r(Z)$ denote the rth semi-invariant of the random variable Z and let $\mu_z^{(r)}$ denote the rth central moment of Z. For $r > 3$ we note

$$\chi_r(Z) = \mu_z^{(r)} + f(\mu_z^{(r-1)}, \cdots, \mu_z^{(2)}), \tag{3.8}$$

where f is a polynomial in $\mu_z^{(r-1)}, \cdots, \mu_z^{(2)}$ each term of which is at least degree 2 (c.f. [1], page 66). Thus $\chi_r(x_{nk}) = \mu_{nk}^{(r)} + f(\mu_{nk}^{(r-1)}, \cdots, \mu_{nk}^{(2)})$. Now if X and Y are independent random variables we note ([1], page 64) $\chi_r(X + Y) = \chi_r(X) + \chi_r(Y)$. Hence since $S_n = x_{n1} + \cdots + x_{nk_n}$ is the sum of independent random variables we see

$$\chi_r(S_n) = \chi_{r(n)} = \sum_{k=1}^{k_n} \mu_{nk}^{(r)} + \left\{ \sum_{k=1}^{k_n} f(\mu_{nk}^{(r-1)}, \cdots, \mu_{nk}^{(2)}) \right\}. \tag{3.9}$$

The general term of the expression in braces may be written as $T = c \sum_{k=1}^{k_n} \prod_{i=1}^{p} \mu_{nk}^{(s_i)}$ where c is a constant, $2 \leqq s_i < r$, $p \geqq 2$ and where $s_i = s_j$ does not imply $i = j$. But

$$|T| \leqq c \max_k |\mu_{nk}^{(s_1)}| \max_k |\mu_{nk}^{(s_2)}| \cdots \max_k |\mu_{nk}^{(s_{p-1})}| \sum_{k=1}^{k_n} |\mu_{nk}^{(s_p)}| .$$

Thus by (3.6) and (3.7) we see that $T \to 0$ as $n \to \infty$. Since the number of terms in f depends only on r this shows that the quantity in braces in (3.9) approaches zero as $n \to \infty$. This proves the Lemma.

Proof of Theorem 6. We note

$$\sum_{k=1}^{k_n} \int_{-\infty}^{\infty} x^r \, dF_{nk}(x + \mu_{nk}) = \int_{-\infty}^{\infty} (x^{r-2} + x^r) \, dG_n(x),$$

where $G_n(x) = \sum_{k=1}^{k_n} \int_{-\infty}^{x} [u^2/(1 + u^2)] \, dF_{nk}(u + \mu_{nk})$ as defined in Lemma 1.

Now by (3.5)

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} x^k \, dG_n(x) = \int_{-\infty}^{\infty} x^k \, dG(x) \quad k = 0, 1, 2, \cdots$$

and therefore for $r \geqq 2$,

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} (x^{r-2} + x^r) \, dG_n(x) = \int_{-\infty}^{\infty} (x^{r-2} + x^r) \, dG(x).$$

But $\int_{-\infty}^{\infty} (x^{r-2} + x^r) \, dG(x)$ is by (3.1) the rth semi-invariant of the infinitely divisible distribution $F(x)$. Thus (for $r > 1$)

$$\lim_{n\to\infty} \sum_{k=1}^{k_n} \int_{-\infty}^{\infty} x^r \, dF_{nk}(x + \mu_{nk}) = \chi_r \equiv r\text{th semi-invariant of } F(x). \tag{3.10}$$

Using (3.10) and Lemma 3 we obtain

$$\lim_{n\to\infty} \chi_{r(n)} = \chi_r, \tag{3.11}$$

that is the rth semi-invariant of $F_n(x)$ approaches the rth semi-invariant of $F(x)$ as $n \to \infty$. Let $\mu^{(k)} = \int_{-\infty}^{\infty} (x - \mu)^k \, dF(x)$. By Lemmas 1 and 2 we have $\mu_n \to \mu$, $\mu_n^{(2)} \to \mu^{(2)}$ as $n \to \infty$. Now $\chi_3 = \mu^{(3)}$, $\chi_{3(n)} = \mu_n^{(3)}$; $\chi_4 = \mu^{(4)} - 3(\mu^{(2)})^2$, $\chi_{4(n)} = \mu_n^{(4)} - 3(\mu_n^{(2)})^2$ and in general as indicated in (3.8) $\chi_r = \mu^{(r)} + f(\mu^{(r-1)}, \cdots, \mu^{(2)})$, where f is a polynomial and $\chi_{r(n)} = \mu_n^{(r)} + f(\mu_n^{(r-1)}, \cdots, \mu_n^{(2)})$. Using (3.11) and an induction argument we see $\lim_{n\to\infty} \mu_n^{(r)} = \mu^{(r)}$, $(r \geqq 2)$ and this together with Lemma 2 completes the proof of Theorem 6.

REFERENCES

[1] B. V. Gnedenko and A. N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*, translation by K. L. Chung, Addison-Wesley, 1954.

[2] J. M. Shapiro, "A condition for existence of moments of infinitely divisible distributions," *Canadian J. Math.*, Vol. 8 (1956), pp. 69–71.

ON THE IDENTITY RELATIONSHIP FOR FRACTIONAL REPLICATES OF THE 2^n SERIES[1]

By R. C. Burton and W. S. Connor

National Bureau of Standards

1. Summary. The paper considers $\frac{1}{2^r}$ fractional replication designs of the factorial series with n factors each at two levels. The identity relationship for such designs is often written in terms of a symbol I and collections of letters which denote interactions among the factors. These collections may conveniently be called "words," and the number of letters in a collection, the "length" of the word. The problem considered is that of the existence of an identity relationship which contains words of specified lengths.

It is known that the words of an identity, together with the symbol I, form an Abelian group. The group contains sets of independent generators, and the products of such generators. Necessary and sufficient conditions are developed for the existence of an identity relationship for which the lengths of a set of independent generators and their products are specified. Further, it is shown how to construct such an identity relationship, and it is proved that the identity relationship is unique, apart from renaming the letters.

For the more general case in which the lengths of the words are given—but are not associated with particular generators and products—a necessary condition is developed for the existence of the identity relationship. It is shown by example that this condition is not sufficient.

2. Introduction. We shall consider the case of n factors each at two levels. Let the factors be denoted by $A, B, \cdots$ and consider a $\frac{1}{2^r}$ fraction of the 2^n factorial design ($n > r$). The identity relationship consists of the symbol I and $2^r - 1$ words, connected by equality signs, as follows:

$$I = A^{a_1}B^{b_1} \cdots = \cdots = A^{a_m}B^{b_m} \cdots ,$$

where $m = 2^r - 1$; $a_i, b_j, \cdots (j = 1, \cdots, m)$ take on the values 0 or 1; and $A^0 = B^0 = \cdots = 1$. We shall consider the case in which all n letters are present in the identity relationship.

The product of any two words,

$$A^{a_x}B^{b_x} \cdots \qquad \text{and} \qquad A^{a_y}B^{b_y} \cdots ,$$

is $A^{a_x+a_y}B^{b_x+b_y} \cdots$, where $a_x + a_y, b_x + b_y, \cdots$ are reduced modulo 2; and the product of any word with I is the word itself. The words of the identity relationship, together with I, form an Abelian group, in which each word is its own inverse.

Received September 4, 1956.

[1] Work performed (in part) under the sponsorship of the Bureau of Ships, Navy Department.

For a $\frac{1}{2^r}$ fraction, the group contains r words which are independent generators. The product of any i ($i = 2, \cdots, r$) of these generators is a word different from all of the generators. To illustrate, consider $n = 5$ and $r = 3$. An identity relationship is

$$I = \underline{ABC} = \underline{CDE} = \underline{AE} = ABDE = BCE = ACD = BD,$$

where a set of independent generators has been underscored.

Before proceeding with our main argument, it is helpful to review some known necessary conditions. For this purpose we shall let $w_1, \cdots, w_n$ denote the lengths of the words. Brownlee, Kelly, and Loraine [1] state that the following conditions are necessary:

(i) $\sum_{i=1}^{n} w_i = 2^{r-1}n$.

(ii) Either the w's all are even or 2^{r-1} of them are odd.

(iii) When 2^{r-1} of the w's are odd, the words with even numbers of letters must, with the identity I, form a subgroup of order 2^{r-1}.

(iv) If some $w = n$, the remaining w's must be divisible into pairs such that the total of each pair is n.

(v) If some $w = 1$, the remaining numbers must be divisible into pairs such that the numbers in each pair differ by 1.

They then state that "no other necessary conditions appear susceptible to expression in simple general terms."

These conditions do not require that the lengths of particular generators and their products be specified. The problem which we shall solve does make this requirement. We shall state necessary and sufficient conditions for the existence of an identity relationship, for which the lengths of the generators and their products are given. In addition, we shall prove that such an identity relationship is unique, apart from the naming of letters; and we shall show how to construct the identity relationship.

For the more general case considered by Brownlee, Kelly, and Loraine, we shall derive an additional necessary condition, and shall show by example that it, together with the above conditions (i), (ii), (iii), is not sufficient to imply the existence of an identity relationship.

In addition to [1], some papers which treat fractional replicates and the related problem of block confounding are listed as references [2], $\cdots$, [7].

3. The existence of an identity relationship with generators and products of specified lengths. Let $i_1, i_2, \cdots, i_s$ be s integers such that $0 < i_1 < i_2 < \cdots < i_s < r + 1$. The ith generator will be denoted by $W(i)$ and the product of the i_1-th i_2-th, $\cdots$ and i_s-th generator by $W(i_1, i_2, \cdots, i_s)$. There are exactly $2^r - 1$ words corresponding to the $2^r - 1$ symbols $(i_1, i_2, \cdots, i_s)$. The numbers of letters in the word $W(i_1 \cdots i_s)$ will be denoted by $w = w(i_1 \cdots i_s)$.

We shall find it convenient to introduce a symbol S to denote the entire collection of $2^r - 1$ symbols $(i_1 \cdots i_s)$, a symbol $O = O(i_1 \cdots i_s)$ to denote the collection of symbols which contain an odd number of indices from $(i_1 \cdots i_s)$, and a symbol $E = E(i_1 \cdots i_s)$ to denote the collection of symbols which contain none or an even number of indices from $(i_1 \cdots i_s)$.

Let $n(O)$ and $n(E)$ denote the numbers of symbols in O and E. It is readily shown that

$$n(O) = 2^{r-1},$$
$$n(E) = 2^{r-1} - 1. \tag{3.1}$$

We note that the distribution of a letter among the words of the identity is determined by its distribution among the generators. Suppose that a letter occurs in the s generators $W(i_1), \cdots, W(i_s)$, but not in the remaining generators. Then it occurs in all products which have an odd number of indices from among $i_1, \cdots, i_s$. But there are $2^r - 1$ ways in which a letter may be distributed among the generators, and hence among the words of the identity.

We shall use the symbol $t = t(i_1 \cdots i_s)$ to denote the number of letters which occur in all of the s generators $W(i_1), \cdots, W(i_s)$ but not in the remaining generators.

For example, consider three generators $W(1) = BCD$, $W(2) = ACDEF$, $W(3) = ACF$. We may illustrate with a Venn diagram which letters are common, denoting the generators by the interiors of the circles in Fig. 1.

Since the t's are the numbers of letters in the basic disjoint sets, it is obvious that

$$\sum_s t(i_1 \cdots i_p) = n. \tag{3.2}$$

It is also clear that any set of t's which are positive integers or zero and satisfy (3.2) corresponds to a constructible identity relationship involving n letters.

We shall now show explicitly how the t's uniquely determine the w's, and conversely the w's uniquely determine the t's.

From the definitions of t, w, and O, it follows that

$$\sum_{o(i_1 \cdots i_s)} t(j_1 \cdots j_p) = w(i_1 \cdots i_s). \tag{3.3}$$

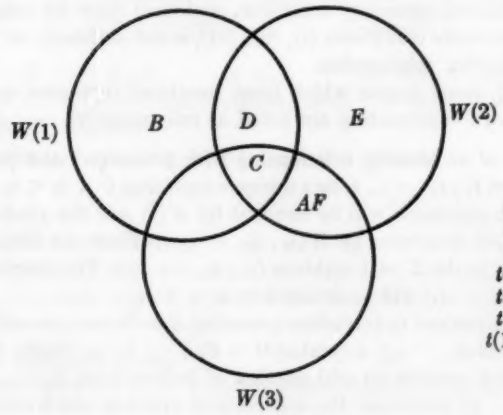

FIG. 1

There are $2^r - 1$ such equations. These equations uniquely determine the w's from the t's.

Let us now introduce a dummy variable $t(0)$, which we define to be identically zero. We may add $t(0)$ to the left-hand side of (3.2) to obtain

$$t(0) + \sum_s t(j_1 \cdots j_p) = n. \tag{3.4}$$

Multiplying (3.3) by 2, and subtracting (3.4) from the product, we obtain

$$-t(0) + \sum_o i(j_1 \cdots j_p) - \sum_E t(j_1 \cdots j_p) = 2w(i_1 \cdots i_s) - n, \tag{3.5}$$

where

$$O = O(i_1 \cdots i_s) \quad \text{and} \quad E = E(i_1 \cdots i_s).$$

We observe that the matrix of coefficients in (3.4) and (3.5) is a Hadamard matrix of order 2^r. Dividing these equations by $2^{r/2}$, the matrix of coefficients becomes an orthogonal matrix.

It is helpful to express the equations in matrix notation. Let t denote the column vector $[t(0), t(1), \cdots, t(12 \cdots r)]$, x the column vector $2^{-r/2}[n, 2w(1) - n, \cdots, 2w(12 \cdots r) - n]$, and C the matrix of coefficients in (3.4) and (3.5) after division by $2^{r/2}$. In this notation the equations may be written as

$$Ct = x. \tag{3.6}$$

Because C is orthogonal, $C^{-1} = C'$ and

$$C^{-1}Ct = C^{-1}x,$$

$$t = C'x. \tag{3.7}$$

The typical equation in (3.7) is

$$\sum_o w(j_1 \cdots j_p) - \sum_E w(j_1 \cdots j_p) = 2^{r-1}t(i_1 \cdots i_s). \tag{3.8}$$

These last equations uniquely determine the t's from the w's. If the w's determine t's which are positive integers or zero, and are such that their sum is n, then the identity relationship can be constructed, and is unique, except for the renaming of letters.

We may sum up in the following theorem.

THEOREM 1. *For a $\frac{1}{2^r}$ fractional replicate of a 2^n factorial design, let the numbers of letters in the words of the identity relationship be $w(1), w(2), \cdots, w(r); w(12), w(13), \cdots, w(r-1, r); \cdots; w(12 \cdots r)$, where $(1), (2), \cdots, (r)$ refer to a set of independent generators and $(12), (13), \cdots (r-1, r); \cdots; (12 \cdots r)$ to their products. With respect to any one of these 2^{r-1} symbols $(i_1 \cdots i_s)$, the remaining symbols divide into a collection $O = O(i_1 \cdots i_s)$ of 2^{r-1} symbols which have an odd number of indices in common with the given symbol, and a collection $E = E(i_1 \cdots i_s)$ of $2^{r-1} - 1$ symbols which have none or an even number of indices in common with the given symbol. If the w's satisfy the $2^r - 1$ equations*

$$\sum_o w(j_1 \cdots j_p) - \sum_E w(j_1 \cdots j_p) = 2^{r-1}t(i_1 \cdots i_s),$$

where $\sum_o t + \sum_e t = n$ *in the sense of implying t's which are positive integers or zero, then the identity relationship exists. Conversely, if the identity relationships exists, then the equations are satisfied. Furthermore, knowledge of the t's is sufficient to construct the identity, and the identity corresponding to a set of t's is unique, apart from renaming the letters.*

4. A necessary condition. We now return to the more general problem of the existence of an identity relationship having words of given length, but without the specification of the lengths of particular generators and their products. We shall develop a necessary condition for the identity relationship to exist.

From (3.6) we have

$$t'C'Ct = x'x, \tag{4.1}$$

and because $C' = C^{-1}$,

$$\begin{aligned} t't &= x'x, \\ \sum w^2 &= 2^{r-2}(\sum t^2 + n^2). \end{aligned} \tag{4.2}$$

Thus, there must exist n or fewer positive integers—the non-zero t's—which add to n and satisfy (4.2).

We may consider the special case in which the variance of the w's is a minimum. The variance of w is defined by

$$\begin{aligned} V(w) &= [\sum w^2 - (\sum w)^2 / (2^r - 1)] / (2^r - 1) \\ &= 2^{r-2}[(2^r - 1)\sum t^2 - n^2] / (2^r - 1)^2. \end{aligned} \tag{4.3}$$

Now $V(w)$ is a minimum when $\sum t^2$ is a minimum—i.e., when $\sum t^2 = n$. In this case (4.2) becomes

$$\sum w^2 = 2^{r-2}n(n + 1). \tag{4.4}$$

We may sum up in a theorem and corollary.

THEOREM 2. *For a* $\frac{1}{2^r}$ *fractional replicate of a* 2^n *factorial design, let the numbers of letters in the words of the identity relationship be* $w_1, \cdots, w_m$, *where* $m = 2^r - 1$. *Then a necessary condition for the identity relationship to exist is that there exist n or fewer positive integers whose sum is n and whose squares add to* $2^{-r+2}\sum w^2 - n^2$.

COROLLARY. *If the variance of w is a minimum, then it is necessary that* $\sum w^2 = 2^{r-2}n(n + 1)$.

We shall show by example that this condition, together with conditions (i), (ii), (iii) are not sufficient conditions. Conditions (iv) and (v) do not apply in the example.

Let $n = 9$ and $r = 4$. Let the distribution of w's be as follows:

w	Frequency
4	7
5	6
7	2

For this distribution, $\sum w = 72$ and $\sum w^2 = 360$. This distribution has minimum variance, and satisfies the corollary. In addition, it satisfies (i), (ii), and (iii). To see that it satisfies (iii), we have only to write the following identity:

$$I = ABCD = ABEF = ACEG = CDEF = BDEG = BCFG = ADFG.$$

We shall show that the necessary conditions of Theorem 1 are not satisfied. To do this, we shall make a unique identification of the w's with the generators and their products.

Let us choose $w(1) = 7$ and $w(2) = 7$. Then because two odd lengths imply an even length, $w(12)$ must be 4. We choose $w(3) = 5$, which implies that $w(13) = w(23) = 4$, and $w(123) = 5$. Finally, we choose $w(4) = 5$, which completely determines the remaining products. The w's are as follows:

$w(1) = 7$	$w(12) = 4$	$w(123) = 5$	$w(1234) = 4$
$w(2) = 7$	$w(13) = 4$	$w(124) = 5$	
$w(3) = 5$	$w(14) = 4$	$w(134) = 5$	
$w(4) = 5$	$w(23) = 4$	$w(234) = 5$	
	$w(24) = 4$		
	$w(34) = 4$		

Now

$$\sum_{O(1)} w - \sum_{E(1)} w = 4$$

which by (3.8) implies that $t(1) = \frac{1}{2}$. Accordingly, the identity relationship does not exist.

REFERENCES

[1] K. A. Brownlee, B. K. Kelly, and P. K. Loraine, "Fractional replication arrangements for factorial experiments with factors at two levels," *Biometrika*, Vol. 35 (1948), pp. 268–276.

[2] O. Kempthorne, "A simple approach to confounding and fractional replication in factorial experiments," *Biometrika*, Vol. 34 (1947), pp. 255–272.

[3] D. J. Finney, "The fractional replication of factorial arrangements," *The Annals of Eugenics, London*, Vol. 12 (1945), pp. 291–301.

[4] R. A. Fisher, "The theory of confounding in factorial experiments in relation to the theory of groups," *Annals of Eugenics, London*, Vol. 11 (1942), pp. 341–353.

[5] C. R. Rao, "Hypercubes of strength 'd' leading to confounded designs in factorial experiments," *Bulletin Calcutta Math. Soc.*, Vol. 38 (1946), pp. 67–68.

[6] R. L. Plackett and J. P. Burman, "The design of optimum multifactorial experiments," *Biometrika*, Vol. 33 (1943–1946), pp. 305–325.

[7] R. C. Bose and K. A. Bush, "Orthogonal arrays of strength two and three," *Annals of Math. Stat.*, Vol. 23 (1952), pp. 508–524.

NOTES

WAITING TIMES WHEN QUEUES ARE IN TANDEM

By Edgar Reich

University of Minnesota

1. We study the distribution of waiting times when customers proceed to a second (multiple-counter) queue after having been processed at a first (multiple-counter) queue[1]. For reasons of expediency we restrict ourselves to the case of unsaturated queues in "equilibrium," that is, to stationary statistics. The main results are for the case of exponential service time, where it turns out that, contrary to a-priori intuition, the situation is surprisingly simple. As shown by Theorem 6, no such simple behavior can be expected when the service time distributions are even only slightly more general. Theorem 4 was first found essentially by P. J. Burke [1], by a different method.[2]

The concept of reversibility of a Markov chain, certain aspects of which are discussed in Sec. 2, has turned out to be fruitful in connection with the analysis, and is of some independent interest.

2. A stationary stochastic process $N(t)$ is said to be reversible if $N(t)$ and $N(-t)$ have the same multivariate distributions. If $N(t)$ is a discrete or continuous parameter Markov chain with a denumerable state space, say, 0, 1, 2, $\cdots$, then $N(-t)$ is a process of the same type. The necessary and sufficient condition for reversibility becomes

$$\theta_{ij}(t) = p_i P_{ij}(t) = p_j P_{ji}(t) = \theta_{ji}(t), \qquad i, j = 0, 1, 2, \cdots, \tag{1}$$

where p_i and $P_{ij}(t)$ are respectively, the stationary, and transition probabilities of $N(t)$.

Kolmogorov's criterion for reversibility of Markov chains with a finite state space ([8]; [5], p. 66) may, in a special case, be immediately generalized to the denumerable state-space case, as follows.

Theorem 1. *Let $N(k)$, $k = 0, \pm 1, \pm 2, \cdots$, be an irreducible stationary discrete-parameter Markov chain with the state space* 0, 1, 2, $\cdots$, *the stationary probabilities u_k, and the singlestep transition probabilities π_{ij}. A necessary and sufficient condition for the reversibility of $N(k)$ is that*

$$\pi_{i_1 i_2}\pi_{i_2 i_3} \cdots \pi_{i_{n-1} i_n}\pi_{i_n i_1} = \pi_{i_1 i_n}\pi_{i_n i_{n-1}} \cdots \pi_{i_3 i_2}\pi_{i_2 i_1} \tag{2}$$

for every sequence of non-negative integers $(i_1, i_2, \cdots, i_n, i_1)$ beginning and ending with the same integer.

Received July 15, 1956; revised November 27, 1956.

[1] A part of this paper represents work done at The RAND Corporation.

[2] A special case of a part of this theorem was also treated (unpublished) by H. H. Goode and R. E. Machol. Their work is to appear in a text.

Proof. According to (1), for $t = 1$,

$$\theta_{i_1 i_2}\theta_{i_2 i_3}\cdots\theta_{i_{n-1}i_n}\theta_{i_n i_1} = \theta_{i_1 i_n}\theta_{i_n i_{n-1}}\cdots\theta_{i_2 i_2}\theta_{i_2 i_1} \tag{3}$$

is necessary. Since $u_i > 0$, we may cancel the u_i, obtaining (2). Summing both sides of (2) over $i_3, i_4, \cdots, i_n$, we find

$$\pi_{i_1 i_2}\pi_{i_2 i_1}(n-1) = \pi_{i_1 i_2}(n-1)\pi_{i_2 i_1}.$$

If τ is the period of the chain, then the $\limsup_{n\to\infty}$ of the left and right sides are $\tau u_{i_1}\pi_{i_1 i_2}$ and $\tau u_{i_2}\pi_{i_2 i_1}$, respectively ([4], p. 331). Hence $u_{i_1}\pi_{i_1 i_2} = u_{i_2}\pi_{i_2 i_1}$. Eq. (1) follows by induction.

Let us call a finite sequence of non-negative integers, beginning and ending in the same integer a *cycle* if no proper portion begins and ends in the same integer. Evidently (2) need hold only for cycles.

Consider a continuous-parameter, time-homogeneous Markov chain $N(t)$, with $P_{ij}(h) = \delta_{ij} + hL_{ij} + o(h)$,

$$\sum_{j=0}^{\infty} L_{ij} = 0,\ L_{ii} < 0,\ 1, 2, \cdots.$$

A process of this type will be said to be of type A if, in addition,

(i) $N(t)$ has stationary probabilities p_i, $\sum_i p_i = 1$, $\sum_i p_i L_{ij} = 0$;

(ii) $N(t)$ has at most a finite number of discontinuities in every finite interval; $-\sum_{i=0}^{\infty} p_i L_{ii} < \infty$.

(iii) the associated discrete-parameter chain N^* defined by

$$\pi_{ij} = (\delta_{ij} - 1)L_{ij} / L_{ii}, \qquad i, j = 0, 1, 2, \cdots,$$

is irreducible. (Note: This is a chain, of period $\geqq 2$, resulting from a shift of the instants at which $N(t)$ changes state to the instants $t = \cdots -1, 0, 1, \cdots$)

THEOREM 2. *A necessary and sufficient condition for a continuous-parameter Markov process of type A to be reversible is that*

$$L_{i_1 i_2}L_{i_2 i_3}\cdots L_{i_{n-1}i_n}L_{i_n i_1} = L_{i_1 i_n}L_{i_n i_{n-1}}\cdots L_{i_2 i_2}L_{i_2 i_1} \tag{4}$$

for every cycle.

Proof. Since the matrix $P_{ij}(t)$ of a process of type A is uniquely determined by its values for infinitesimal t, condition (1) is equivalent to

$$p_i L_{ij} = p_j L_{ji}. \tag{5}$$

Note that

$$r = \left(-\sum_{i=0}^{\infty} p_i L_{ii}\right)^{-1} > 0,$$

in view of the second part of assumption (ii), above.

$$u_i = -r p_i L_{ii}, \qquad i = 0, 1, \cdots,$$

can be assigned as stationary probabilities to N^*. Then a necessary and sufficient condition for N^* to be reversible is

$$u_i\pi_{ij} = -rp_i(\delta_{ij} - 1)L_{ij} = -rp_j(\delta_{ji} - 1)L_{ji},$$

which is equivalent to (5). The theorem follows from the fact that (4) is equivalent to (2).

Let $B(t)$ denote a stationary birth-death process with state space $0, 1, 2, \cdots$, the stationary probabilities p_i, and

$$P_{i,i+1}(h) = \lambda_i h + o(h), \qquad \lambda_i > 0, \qquad i = 0, 1, \cdots,$$

$$P_{i,i-1}(h) = \mu_i h + o(h), \qquad \mu_i > 0, \quad i = 1, 2, \cdots, \mu_0 = 0,$$

$$P_{ii}(h) = 1 - \lambda_i h - \mu_i h + o(h), \qquad i = 0, 1, \cdots.$$

$B(t)$ is permitted to have at most a finite number of increases (births), and decreases (deaths) in any finite time interval, $\sum p_i(\lambda_i + \mu_i) < \infty$.

THEOREM 3. *$B(t)$ is reversible.*

Proof. If (4) is to be other than of the form $0 = 0$, the cycle must be of length 3, in which case (4) still holds trivially[3].

COROLLARY. *If $\lambda_n = \lambda$, $(n = 0, 1, \cdots)$ then the death times of $B(t)$ form a Poisson process of density λ.*

Outline of proof. Since λ_n is constant the birth times are Poisson with density λ. The stochastic process $B_1(t) = B(-t)$ is statistically identical with the process $B(t)$. But if $B(t)$ is a fixed realization, and $B_1(t) = B(-t)$ then the births of $B(t)$ become the deaths of $B_1(t)$.

3. Consider an unsaturated queue of type $M/M/s$ (Poisson input, s counters, exponential service time, first come, first served), in equilibrium. If $n(t)$ is the sum of the number of customers on queue, plus those being served, then $n(t)$ is a process of type $B(t)$, in which customers' arrivals correspond to births, and departures to deaths. By considering the reversibility of $n(t)$, guaranteed by the corollary to Theorem 3, the following is now clear:

THEOREM 4. (a) *The sequence of departure times form a Poisson process.* (b) *The value of $n(t)$ is independent of all past departure times.* (c) *If t_0 is a departure time, then $n(t_0 + 0)$ is independent of all past departure times.*

Note. The above results are, of course, true for more general queue disciplines. The number of servers, instead of being fixed, may be permitted to vary as a specified function of the number of customers present. Also, instead of "first come, first served," e.g., random service, or "last come, first served," will do without effect on the results.

Suppose the customers, after departing from a first queue of type $M/M/s$,

[3] Heuristic forms of the necessary argument date to P. and T. Ehrenfest [3]. The above proof is in the spirit of the Ehrenfests' reasoning. Simple algebraic verifications are also possible, but they leave the situation less lucid. The condition $\Sigma p_i(\lambda_i + \mu_i) < \infty$ is actually superfluous.

enter a second multiple-counter queue, where they are served first come, first served, with exponential service time. Such a combination of two tandem queues will be referred to as a σ-system. It follows, from Theorem 4b, that if $n_1(t)$, $n_2(t)$ refer, respectively, to the first and second queues of a σ-system, then $n_1(t)$ and $n_2(\tau)$ are independent, $\tau \leqq t$. This was first proved in the special case $s = 1$, $t = \tau$, by Jackson [6].

In what follows, the term *waiting time* will be used to refer to the time elapsed between a customer's arrival and departure, the service time included. Let T_1 and T_2 denote a customer's waiting time at the first and second queues of a σ-system, respectively.

THEOREM 5. *If $s = 1$, then T_1 and T_2 are independent.*

Proof. Let n_1 be the number of customers at the first queue the instant after a customer C departs, and let n_2 be the number of customers C finds at the second queue (customers being served included). As a corollary of Theorem 4c, n_1 and n_2 are independent. Let

$$A(t; k) = \Pr\{T_1 < t \mid n_2 = k\}.$$

If λ is the number of customers arriving per unit time, then n_1 is the number of Poisson events of density λ that occurred during the waiting period T_1. We have

$$\Pr\{n_1 = j \mid T_1 = t, n_2 = k\} = e^{-\lambda t}(\lambda t)^j/j!.$$

Therefore

$$E\{z^{n_1} \mid n_2 = k\} = \int_0^\infty e^{\lambda t z} e^{-\lambda t}\, dA(t; k).$$

Now the left side is independent of k. Therefore $A(t; k)$ does not depend on k. Hence n_2, and consequently also T_2, are independent of T_1.

4. We will now consider the queues of type $E_j / E_k / s$ (interarrival and service periods normalized chi-square with $2j$ and $2k$ degrees of freedom, respectively [7]), and show that Theorem 4a cannot be generalized further, in a certain direction. Note that both when $j = k = 1$, and $j = k = \infty$, the departure epochs of an $E_j / E_k / s$ queue are again E_k. (The case $j = k = \infty$ corresponds to a periodic input with constant service time.) One may therefore ask[4] whether this state of affairs holds whenever $j = k$. However, Theorem 6, below, shows this to be false.

THEOREM 6. *The departure epochs of an $E_2 / E_2 / 1$ process are not an E_2 process.*

Proof. For the case under consideration we have x = interarrival period = $x_1 + x_2$, where x_1, $i = 1, 2$, are independent, with

$$E\{e^{-sx_i}\} = \frac{\lambda}{\lambda + s}, \qquad \lambda > 0, \qquad i = 1, 2.$$

[4] This question is related to the asymptotic behavior of a large number of queues in tandem, each with E_k-type service time, k fixed.

Similarly, the service periods, y, are of the form $y = y_1 + y_2$, where y_i, $i = 1, 2$, are independent, and

$$E\{e^{-sy_i}\} = \frac{\mu}{\mu + s}, \qquad 0 < \rho = \lambda/\mu < 1, \qquad i = 1, 2.$$

At a given instant the entrance will be said to be in state 1 (2) if the system is in the portion $x_1(x_2)$ of an interarrival period. Consider instants just following a departure. Let $A_{i0} = \Pr\{\text{extrance is in state } i, \text{ and there are 0 customers left behind}\}$, $i = 1, 2$. Let τ be the length of an interdeparture period. If the departure epochs formed an E_2 process it would follow that τ had the same marginal distribution as x, that is,

$$E\{e^{-s\tau}\} = A_{10}\left(\frac{\lambda}{\lambda + s}\right)^2\left(\frac{\mu}{\mu + s}\right)^2 + A_{20}\left(\frac{\lambda}{\lambda + s}\right)\left(\frac{\mu}{\mu + s}\right)^2 + (1 - A_{10} - A_{20})\left(\frac{\mu}{\mu + s}\right)^2 = \left(\frac{\lambda}{\lambda + s}\right)^2.$$

Multiplying both sides by $(\lambda + s)^2(\mu + s)^2$, and equating coefficients of s^2, we have

$$P_0 = \Pr\{0 \text{ customers are left behind by a departing customer}\} = A_{10} + A_{20} = 1 - \rho^2.$$

However this is incorrect, as it differs from Volberg's [9] formula for P_0. Thus we have a contradiction.

A related question is that of the possibility of *imbedding* $n(t)$ in a reversible Markov process, e.g., for $s = 1$. To this end we define the "pseudostate" $\tilde{n}(t)$ of an $E_j / E_k / 1$ queue. We shall say that $a(t) = r$, $r = 0, 1, 2, \cdots, j - 1$, if the $(r + 1)$st stage of the interarrival period is in progress. Similarly, put $b(t) = r$, $r = 0, 1, \cdots, k - 1$, if the $(r + 1)$st stage of the service period is in progress; if the counter is empty, $b(t) = 0$. Define

$$\tilde{n}(t) = n(t) + \frac{a(t)}{j} - \frac{b(t)}{k}. \tag{6}$$

The realizations of the process $\tilde{n}(t)$ are constant except for jumps of height $1/j$, upward, and jumps of height $1/k$, downward. We make the following observation.

THEOREM 7[5]. *If j and k are relatively prime, $\tilde{n}(t)$ is a Markov process.*

Proof. If the hypothesis is satisfied, $n(t_0)$, $a(t_0)$, $b(t_0)$ can be recovered from a knowledge of $\tilde{n}(t_0)$.

We conclude that if j and k are relatively prime, and $j = k$, then $\tilde{n}(t)$ is reversible. Since $j = k = 1$ is the only admissible possibility, the special nature of the $E_1 / E_1 / 1$ queue is seen in a new light.

A straightforward computation shows that the following partial *converse* of Theorem 4a holds.

[5] This fact enables one to study the transient behavior of $n(t)$ for $E_j/E_k/1$, j, k relatively prime. We shall not explore this further at this time, however.

THEOREM 8. *If the arrival and departure epochs of a single-counter queue are both Poisson, then the service time distribution is exponential, or a step function at* 0.

The author has had valuable discussions with A. W. Marshall and T. E. Harris in connection with this work.

REFERENCES

[1] P. J. BURKE, "The Output of a Queuing system," *Operations Research*, Vol. 4 (1956), pp. 699–704.
[2] J. L. DOOB, *Stochastic Processes*, John Wiley and Sons, New York, 1953.
[3] P. EHRENFEST AND T. EHRENFEST, "Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem," *Physikalische Zeitschrift*, Vol. 8 (1907), pp. 311–314.
[4] W. FELLER, *An Introduction to Probability Theory and its Applications*, Vol. 1, John Wiley and Sons, New York, 1950.
[5] M. FRÉCHET, *Méthode des fonctions arbitraires. Théorie des événements en chaîne dans le cas d'un nombre fini d'etats possibles*, Gauthier-Villars, Paris, 1938.
[6] R. R. P. JACKSON, "Queueing systems with phase type service," *Operational Research Quarterly*, Vol. 5 (1954), pp. 109–120.
[7] D. G. KENDALL, "Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain," *Ann. Math. Stat.*, Vol. 24 (1953), pp. 338–354.
[8] A. KOLMOGOROFF, "Zur Theorie der Markoffschen Ketten," *Math. Ann.*, Vol. 112 (1936), pp. 155–160.
[9] O. A. VOLBERG, "Problème de la queue stationnaire et nonstationnaire," *Doklady Akad. Nauk SSSR*, N.S., Vol. 24 (1939), pp. 657–661.
[10] R. R. P. JACKSON, "Random queueing processes with phase-type service," *J. Roy. Stat. Soc. B*, Vol. 18 (1956), pp. 129–132.

ON THE POWER OF OPTIMUM TOLERANCE REGIONS WHEN SAMPLING FROM NORMAL DISTRIBUTIONS[1]

BY IRWIN GUTTMAN

University of Alberta

1. Introduction and Summary. In [1], optimum β-expectation tolerance regions were found by reducing the problem to that of solving an equivalent hypothesis testing problem. The regions produced when sampling from a k-variate normal distribution were found to be of similar β-expectation and optimum in the sense of minimax and most stringency. It is the purpose of this paper to discuss the "Power" or "Merit" of such regions, when sampling from the k-variate normal distribution.

Let $X = (X_1, \cdots, X_n)$ be a random sample point in n dimensions, where each X_i is an independent observation, distributed by $N(\mu, \sigma^2)$. It is often desirable to estimate on the basis of such a sample point a region which contains a given fraction β of the parent distribution. We usually seek to estimate the center $100\beta\%$ of the parent distribution and/or the $100\beta\%$ left-hand tail of the parent distribution.

Received April 20, 1956; revised December 17, 1956.
[1] Research supported by the University of Alberta General Research Fund.

2. Formulation of the Power Function. Suppose sampling from $N(\mu, \sigma^2)$, where *Case* I: μ, σ^2 *unknown*. For this case the solution of the equivalent hypothesis testing problem (as formulated on p. 171 of [1]) is given by

$$\phi_y(x_1, \cdots, x_n) = 1 \quad \text{if} \quad |W| \leqq a_\beta,$$
$$= 0 \quad \text{if} \quad |W| > a_\beta,$$

where

$$W = \frac{y - \bar{x}}{s_x}, \qquad \bar{x} = n^{-1} \sum x_i, \qquad s_x^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

and $\phi_y(x_1, \cdots, x_n)$ is the characteristic function of the minimax most stringent tolerance region $S(X_1, \cdots, X_n)$. The a_β are constants chosen to give

$$S(x_1, \cdots, x_n) = [\bar{x} - a_\beta s_x, \bar{x} + a_\beta s_x]$$

size β, and are tabulated in Table I of [1].

The power of ϕ (as defined on p. 170 of [1]) and hence of S is determined by the distribution of W under the alternative of the equivalent hypothesis testing problem. That is, we have

$$\text{Power} = P_{\text{Alt.}}(|W| \leqq a_\beta),$$
$$= P_{\text{Alt.}}\left(\frac{|W|}{(\alpha^2 + n^{-1})} \leqq \frac{a_\beta}{(\alpha^2 + n^{-1})}\right). \tag{2.1}$$

Now, under the alternative, $y - \bar{x}$ has variance $(\alpha^2 + n^{-1})\sigma^2$. Thus, $W / (\alpha^2 + n^{-1})^{\frac{1}{2}}$ under the alternative, is the Student's "T" variable with $(n-1)$ degrees of freedom. Hence

$$\text{Power} = P\left(|T| \leqq \frac{a_\beta}{(\alpha^2 + n^{-1})^{1/2}}\right). \tag{2.2}$$

The power measures the "degree of confidence" we have in $S(X_1, \cdots, X_n)$ of covering the centre $100\beta\%$ of $N(\mu, \sigma^2)$, when the "desirability" of covering the centre $100\beta\%$ set is given by

$$Q_{\mu,\sigma^2}(S) = \int_\alpha dN(\mu, \alpha^2\sigma^2), \qquad 0 < \alpha < 1.$$

For example, if it is 99.1% desirable to cover the 95% center part of $N(\mu, \sigma^2)$, then $\alpha = \frac{3}{4}$, and the power is found by (2.2) using $\alpha = \frac{3}{4}$. Values of the power for the regions S, where the desirability of the $100\beta\%$ sets are .99, are given for $\beta = .75, .90, .95$ and $.975$, in Table I.

As an example, consider forming S on the basis of a sample of 7. Then from the tables in [1],

$$S = [\bar{x} - 2.616\, s_x, \bar{x} + 2.616\, s_x].$$

Now, suppose we wish to have 99% confidence that $S(x_1, \cdots, x_7)$ contains 95% of $N(\mu, \sigma^2)$. Then, the "confidence" that S covers 95% of the parent dis-

TABLE I

Power of β-expectation tolerance regions,
$[\bar{x} - a_\beta s_x , \bar{x} + a_\beta s_x]$

	Measure of Desirability = .99			
α	.870167	.760906	.638572	.446594
n \ β	.975	.95	.90	.75
2	.9759	.9545	.9141	.7931
3	.9776	.9618	.9361	.8564
4	.9792	.9682	.9516	.9011
5	.9806	.9733	.9576	.9154
7	.9826	.9772	.9669	.9381
11	.9848	.9809	.9763	.9586
21	.9871	.9847	.9818	.9756
31	.9880	.9863	.9842	.9795
41	.9885	.9872	.9855	.9817
61	.9890	.9881	.9869	.9842
121	.9895	.9890	.9884	.9869

tribution, that is the power of S, is found by entering Table I for $n = 7$ in the $\beta = .95$ column. The power is found to be .9772. That is, if $X_1 , \cdots , X_7 , Y$ are independent normally distributed chance variables, $X_1 , \cdots , X_7$ having identical distributions with mean μ and standard deviation σ, Y having mean μ and standard deviation $\alpha\sigma$, then

$$\Pr(\bar{X} - 2.616\, s_x < Y < \bar{X} + 2.616\, s_x) = \begin{cases} .95 \text{ if } \alpha = 1 \\ .9772 \text{ if } \alpha = .760906. \end{cases}$$

Case II. *Mean unknown, variance known.* The minimax and most stringent tolerance region $S(X_1 , \cdots , X_n)$ of similar β-expectation is given by

$$S(x_1 , \cdots , x_n) = [\bar{x} - b_\beta \sigma , \bar{x} + b_\beta \sigma],$$

where σ^2 is the known value of the variance, and b_β are constants chosen to give S size β. Using the same procedure as for Case I, we have

$$\text{Power} = P\left(|Z| \leq \frac{b_\beta}{(\alpha^2 + n^{-1})}\right), \tag{2.3}$$

where Z is the standard normal variate. Values of the power for this case are given in Table II.

Case III. *Mean known, variance unknown.* The minimax and most stringent tolerance region $S(X_1 , \cdots , X_n)$ of similar β-expectation is

$$S(x_1 , \cdots , x_n) = [\mu - t_{(1-\beta)/2}\, s'_x , \quad \mu + t_{(1-\beta)/2}\, s'_x],$$

TABLE II

Power of β-expectation tolerance regions,
$[\bar{x} - b_\beta\sigma, \bar{x} + b_\beta\sigma]$

	Measure of Desirability = .99			
α	.870167	.760906	.638572	.446594
β / n	.975	.95	.90	.75
2	.9856	.9792	.9655	.9079
3	.9868	.9822	.9726	.9312
4	.9875	.9839	.9766	.9449
5	.9879	.9850	.9792	.9538
7	.9885	.9863	.9822	.9644
11	.9890	.9876	.9850	.9742
21	.9894	.9887	.9874	.9822
31	.9896	.9891	.9882	.9848
41	.9897	.9893	.9887	.9861
61	.9898	.9896	.9891	.9874
121	.9899	.9898	.9896	.9887

TABLE III

Power of β-expectation tolerance regions,
$[\mu - t_{(1-\beta)/2}s'_x, \mu + t_{(1-\beta)/2}s'_x]$, where $s'_x = \sqrt{n^{-1}\sum(x_i - \mu)^2}$

	Measure of Desirability = .99			
α	.870167	.760906	.638572	.446594
β / n	.975	.95	.90	.75
1	.9765	.9578	.9280	.8651
2	.9787	.9678	.9535	.9243
3	.9806	.9751	.9626	.9501
4	.9821	.9771	.9695	.9581
5	.9832	.9787	.9747	.9644
7	.9847	.9811	.9779	.9733
11	.9863	.9838	.9814	.9787
21	.9879	.9865	.9851	.9835
40	.9889	.9881	.9873	.9864
60	.9892	.9887	.9882	.9875
120	.9896	.9893	.9891	.9887

where μ is the known value of the mean, t_α is the point exceeded with probability α by the Student's "T" variable with n degrees of freedom. Using a similar procedure as above, the power of these regions are clearly given by

$$\text{Power} = P(|T| \leqq t_{(1-\beta)/2}/\alpha),$$

where T is the Student's variable with n degrees of freedom. Values of this power are given in Table III. s_x' is defined as $\sqrt{n^{-1}\sum_1^n (x_i - \mu)^2}$.

3. Sampling from a k-variate normal distribution. Consider the case of sampling from a multivariate normal distribution

$$c \exp[-\tfrac{1}{2}(\omega - \mu)\Lambda(\omega - \mu)'],$$

where $\mu = (\mu_1, \cdots, \mu_k)$ and Λ^{-1} is the variance covariance matrix of $\omega = (X_1, \cdots, X_k)$. Suppose μ and Λ^{-1} are unknown, and suppose it is desired to form regions S which cover the centre part of the parent distribution. The choice of the measure of desirability is

$$Q_{\mu,\Lambda^{-1}} = N(\mu, \alpha^2\Lambda^{-1}), \qquad 0 < \alpha < 1. \tag{3.1}$$

It was shown in [1] that the solution of forming these regions, that is, of an equivalent hypothesis testing problem (see p. 176 of [1]) is

$$\begin{aligned} \phi_\beta(\omega_1, \cdots, \omega_n) &= 1 \quad \text{when} \quad (\xi - \bar{\omega})A^{-1}(\xi - \omega)' \leqq c_\beta, \\ &= 0 \quad \text{when} \quad (\xi - \bar{\omega})A^{-1}(\xi - \bar{\omega})' > c_\beta, \end{aligned} \tag{3.2}$$

where $\bar{\omega} = n^{-1}\sum_{\alpha=1}^n \omega_\alpha$,

$$A = (n-1)^{-1}\sum_{\alpha=1}^n (\omega_\alpha - \bar{\omega})'(\omega_\alpha - \bar{\omega}),$$

and c_β are constants chosen to give the ellipsoidal region

$$S(\omega_1, \cdots, \omega_n) = \{\bar{\xi} \mid (\bar{\xi} - \bar{\omega})A^{-1}(\bar{\xi} - \bar{\omega})' \leqq c_\beta\} \tag{3.3}$$

size β, that is, β-expectation. These regions were found to be minimax and most stringent. Letting $\gamma^2 = (\xi - \bar{\omega})A^{-1}(\xi - \bar{\omega})'$, the power clearly takes the form

$$\begin{aligned} \text{Power} &= P_{\text{Alt.}}\{\gamma^2 \leqq c_\beta\} = P_{\text{Alt.}}\left\{\frac{\gamma^2}{\alpha^2 + n^{-1}} \leqq \frac{c_\beta}{\alpha^2 + n^{-1}}\right\} \\ &= P\left\{T^2 \leqq \frac{c_\beta}{\alpha^2 + n^{-1}}\right\}, \end{aligned} \tag{3.4}$$

where T^2 is Hotelling's T^2 variable with $(n-1)$ degrees of freedom, and Alt. refers to the Alternative hypothesis in the formulation of p. 176 of [1].

By making the transformation

$$T^2 = (n-1)\frac{k}{n-k}F,$$

it is well known that Hotelling's T^2 distribution goes into Fisher's F distribution with $k, n-k$ degrees of freedom. That is, the power of S is given by

$$\text{Power} = P\left(F \leqq \frac{n-k}{k(n-1)}\frac{c_\beta}{(\alpha^2 + n^{-1})}\right).$$

TABLE IV

Power of β-expectation tolerance regions,
$(\xi - \bar{\omega})A^{-1}(\xi - \bar{\omega})' \leqq c_\beta$

	Measure of Desirability = .99			
α	.88927	.79697	.69432	.53403
n \ β	.925	.95	.90	.75
3	.9755	.9531	.9105	.7810
4	.9770	.9598	.9318	.8502
5	.9784	.9661	.9522	.9022
7	.9809	.9752	.9606	.9291
11	.9838	.9794	.9751	.9578
21	.9869	.9845	.9818	.9770
30	.9880	.9865	.9847	.9812
31	.9881	.9866	.9849	.9815
32	.9882	.9868	.9851	.9818

In [1], it was shown that $c_\beta = (1 + n^{-1}) \cdot (n - 1) \cdot (k / n - k) \cdot F_{1-\beta}$, where $F_{1-\beta}$ is the point exceeded with probability $1 - \beta$ using the F distribution with k, $n - k$ degrees of freedom. Hence the regions (3.3) have power given by

$$\text{Power} = P\left(F \leqq \frac{1 + n^{-1}}{\alpha^2 + n^{-1}} F_{1-\beta}\right). \tag{3.5}$$

Values of the power function (3.5) are given for the case of sampling from the bi-variate normal distribution ($k = 2$), when the correlation coefficient ρ is zero, and desirability of the centre $100\beta\%$ sets is .99, in Table IV.

REFERENCES

[1] D. A. S. FRASER AND IRWIN GUTTMAN, "Tolerance regions," *Ann. Math. Stat.*, Vol. 27 (1956), p. 162.

[2] H. HOTELLING, "The generalization of Student's ratio," *Ann. Math. Stat.*, Vol. 2 (1931), p. 360.

[3] E. L. LEHMANN, Lectures on "Theory of Testing Hypothesis," recorded by Colin Blyth (1950), University of California Press.

THE CONVERGENCE OF CERTAIN FUNCTIONS OF SAMPLE SPACINGS[1]

BY LIONEL WEISS

Cornell University

1. Introduction and summary. Suppose $g(u_1, \cdots, u_k)$ is a continuous function of its arguments, homogeneous of order r, monotonic nondecreasing in each of its

Received July 23, 1956.

[1] Research under contract with the Office of Naval Research. It may be reproduced in whole or in part for any purpose of the United States Government.

arguments, which is positive whenever each of its arguments is positive, and is such that for any given $K(0 < K < \infty)$, there is a number $R(K)(0 < R(K) < \infty)$ such that $g(u_1, \cdots, u_k) < K$ and $u_1 \geqq 0, \cdots, u_k \geqq 0$ imply that $u_1 + \cdots + u_k < R(K)$.

Let $U_1, \cdots, U_k$ be chance variables with joint density $e^{-(u_1+\cdots+u_k)}$ for $u_1 \geqq 0, \cdots, u_k \geqq 0$, and zero elsewhere. For any t, define $U(t)$ as $P[g(U_1, \cdots, U_k) \leqq t]$. We note that $U(t)$ is a continuous distribution function, with $U(0) = 0$.

Let $\rho(v)$ be a bounded nonnegative function with a finite number of discontinuities, defined for $0 \leqq v \leqq 1$. Suppose $X_1, X_2, \cdots, X_n$ are independently and identically distributed chance variables, each with density $f(x)$, $f(x)$ being bounded, and having a finite number of discontinuities and oscillations. $F(x)$ denotes $\int_{-\infty}^{x} f(x)\,dx$. Define $Y_1 \leqq Y_2 \leqq \cdots \leqq Y_n$ as the ordered values of $X_1, \cdots, X_n$, and define T_i as $Y_{i+1} - Y_i (i = 1, \cdots, n-1)$. Let $R_n(t)$ denote the proportion of the values

$$\rho\left(\frac{1}{n}\right) g(T_1, \cdots, T_k), \qquad \rho\left(\frac{2}{n}\right) g(T_2, \cdots, T_{k+1}), \cdots,$$

$$\rho\left(\frac{n-k}{n}\right) g(T_{n-k}, \cdots, T_{n-1})$$

which are less than or equal to t/n^r.

Let $\overline{U}[[tf^r(x)] / \{\rho[F(x)]\}]$ be defined as follows. If $f(x) = 0$,

$$\overline{U}[[tf^r(x)] / \{\rho[F(x)]\}] = 0$$

regardless of the value of t. If x is such that $f(x) > 0$ and $\rho[F(x)] = 0$, then $\overline{U}[[tf^r(x)] / \{\rho[F(x)]\}] = 1$ regardless of the value of t. If $f(x) > 0$ and $\rho[F(x)] > 0$, then $\overline{U}[[tf^r(x)] / \{\rho[F(x)]\}] = U[[tf^r(x)] / \{\rho[F(x)]\}]$. Let $S(t)$ denote

$$\int_{-\infty}^{\infty} \overline{U}[[t \cdot f^r(x)] / \{\rho[F(x)]\}] f(x)\,dx,$$

and let $V(n)$ denote $\sup_{t \geqq 0} |R_n(t) - S(t)|$. Then $V(n)$ converges to zero stochastically as n increases. This generalizes the result of [1], where $k = 1$, $g(u_1) = u_1$, $\rho(v) = 1$. The present result may be used to construct tests of fit in the presence of unknown location and scale parameters.

2. Proof of the convergence of $V(n)$.

LEMMA 1. *If for each given positive t, $R_n(t)$ converges to $S(t)$ stochastically as n increases, then $V(n)$ converges to zero stochastically as n increases.*

Proof. $S(t)$ is continuous for all $t > 0$, and is continuous on the right at $t = 0$. $S(0+) = \int_{\rho[F(x)]=0} f(x)\,dx$. It is easily seen that $R_n(0)$ converges to $\int_{\rho[F(x)]=0} f(x)\,dx$ with probability one as n increases. The rest of the proof of the lemma is almost exactly the same as the proof of Lemma 1 of [1].

LEMMA 2. *Let $X_1, X_2, \cdots, X_n$ be independent chance variables, each with a uniform distribution on* [0, 1]. *Let M denote the number of these variables falling in the closed interval $[C, D]$, where $0 \leqq C < D \leqq 1$, and let $Y_1 \leqq Y_2 \leqq \cdots \leqq Y_M$ denote the ordered values of the variables in $[C, D]$. Define $W_1 = Y_2 - Y_1, \cdots,$*

$W_{M-1} = Y_M - Y_{M-1}$. *For a given positive* t, *define* $L(n, t)$ *as the total number of values of* $g(W_1, \cdots, W_k)$, $g(W_2, \cdots, W_{k+1})$, $\cdots$, $g(W_{M-k}, \cdots, W_{M-1})$ *which are not greater than* t/n^r. *Then* $[L(n, t)] / n$ *converges to* $(D - C)U(t)$ *stochastically as* n *increases.*

Proof. Define Z_i to be one if $g(W_i, \cdots, W_{k-1+i}) \leqq t/n^r$, and zero otherwise. M/n converges to $(D - C)$ with probability one as n increases. The conditional distribution given M of $Q_1 = MW_{i_1}$,

$$Q_2 = MW_{i_2}, \cdots, Q_L = MW_{i_L} (1 \leqq i_1 < i_2 < \cdots < i_L \leqq M - 1)$$

is easily verified to be

$$\left[D - C - \frac{(q_1 + \cdots + q_L)}{M}\right]^{M-L} \cdot \frac{M!}{M^L(D - C)^M(M - L)!}$$

for $q_1 + \cdots + q_L \leqq M(D - C)$, and zero elsewhere. As M increases, this density approaches $[1/(D - C)^L] \exp\{-(q_1 + \cdots + q_L)/(D - C)\}$ uniformly in any region where $q_1 + \cdots + q_L < K < \infty$. We note that under this limiting density, $Q_1, \cdots, Q_L$ are independent. To say that $g(W_i, \cdots, W_{k-1+i}) \leqq t/n^r$ is the same as saying that $g((n / M)MW_i, \cdots, (n / M)MW_{k-1+i}) \leqq t$, and as n increases the probability of this last occurrence approaches the probability that $g[(MW_i) / (D - C), \cdots, (MW_{k-1+i}) / (D - C)] \leqq t$. Since M approaches infinity with probability one as n increases, and from the restrictions on $g(u_1, \cdots, u_k)$ given in Sec. 1, we can use the limiting distribution of $MW_i, \cdots, MW_{k-1+i}$ to compute the limiting

$$P\{g[(MW_i) / (D - C), \cdots, (MW_{k-1+i}) / (D - C)] \leqq t\},$$

and we get $U(t)$ as this limiting probability.

$$\frac{L(n, t)}{n} = \frac{Z_1 + \cdots + Z_{M-k}}{n} = \frac{M}{n}\left[\frac{Z_1 + \cdots + Z_{M-k}}{M}\right],$$

and from the considerations above, it is easily seen that $E\{L(n, t)/n\}$ approaches $(D - C)U(t)$ as n increases.

Next we show that

$$E\left\{\left[\frac{L(n, t)}{n} - \frac{EL(n, t)}{n}\right]^2\right\}$$

approaches zero as n increases, which will complete the proof of Lemma 2. The expectation in question is equal to

$$(1 / n^2)E\{\textstyle\sum_1^{M-k}(Z_i - EZ_i)^2\} + (1 / n^2)E\{\textstyle\sum\sum_{i \neq j}(Z_i - EZ_i)(Z_j - EZ_j)\}.$$

$\{Z_i\}$ are uniformly bounded variables, and $M - k < n$, therefore the first term in this last expression certainly approaches zero as n increases. As for $\sum\sum_{i \neq j}(Z_i - EZ_i)(Z_j - EZ_j)$, any such term with $|i - j| > k$ has Z_i and Z_j defined in terms of entirely different and nonoverlapping sets of W's, and by the result on the independence of Q's derived above, if $|i - j| > k$, $E(Z_i - EZ_i)\cdot$

$(Z_j - EZ_j)$ must approach zero as n increases. But the number of terms $E(Z_i - EZ_i)(Z_j - EZ_j)$ with $|i - j| \leqq k$ is less than $2kn$. From these considerations, it follows easily that

$$E\left\{\left[\frac{L(n, t)}{n} - \frac{EL(n, t)}{n}\right]^2\right\}$$

approaches zero as n increases.

Now we turn to the proof of the stochastic convergence of $V(n)$. For simplicity, we assume that both $f(x)$ and $\rho(v)$ are continuous, for the time being. Given any positive ϵ, we can find H intervals $I_1 = (-\infty, z_1)$, $I_2 = (z_1, z_2)$, $I_3 = (z_2, z_3), \cdots, I_H = (z_{H-1}, \infty)$, such that the variation of $f(x)$ and of $\rho[F(x)]$ in each of these intervals is less than ϵ. Denote $\inf_{x \text{ in } I_i} \{f(x)\}$ by g_i, $\sup_{x \text{ in } I_i} \{f(x)\}$ by G_i, $\inf_{x \text{ in } I_i} \{\rho[F(x)]\}$ by h_i, $\sup_{x \text{ in } I_i} \{\rho[F(x)]\}$ by H_i. Let M_i denote the number of variables $X_1, X_2, \cdots, X_n$ that fall in I_i. Define $L_i(n, t)$ in terms of the M_i variables falling in I_i just as $L(n, t)$ was defined in terms of the variables falling in $[C, D]$ in Lemma 2. Define $L_i'(n, t)$ in the same way, except that each variable X_i is replaced by $F(X_i)$. Since $F(X_i)$ has a uniform distribution, Lemma 2 states that $[L_i'(n, t)]/n$ converges stochastically to $[F(z_i) - F(z_{i-1})]U(t)$ as n increases, where z_0 denotes $-\infty$, z_H denotes ∞. Also, since $F(Y_{i+1}) - F(Y_i) = f(\theta)[Y_{i+1} - Y_i]$, $Y_i \leqq \theta \leqq Y_{i+1}$, and from the assumptions about $g(u_1, \cdots, u_k)$, we have $L_i'(n, g_i't) \leqq L_i(n, t) \leqq L_i'(n, G_i't)$. $(M_1 + \cdots + M_i)/n$ converges to $F(z_i)$ with probability one as n increases, therefore the probability approaches one that

$$\frac{\sum_1^H L_i\left(n, \frac{t}{H_i}\right)}{n} - \frac{2H}{n} \leqq R_n(t) \leqq \frac{\sum_1^H L_i\left(n, \frac{t}{h_i}\right)}{n} + \frac{2H}{n}.$$

This implies that the probability approaches one that

$$\sum_1^H [F(z_i) - F(z_{i-1})]U\left(\frac{g_i't}{H_i}\right) \leqq R_n(t) \leqq \sum_1^H [F(z_i) - F(z_{i-1})]U\left(\frac{G_i't}{h_i}\right).$$

But by taking ϵ small enough (i.e., increasing H properly) the two extremes of this last inequality approach $S(t)$, proving the stochastic convergence of $V(n)$.

In the case where $\rho(v)$ and/or $f(x)$ have discontinuities, we can enclose the points of discontinuity in intervals whose total probability is arbitrarily small, change $\rho[F(x)]$ and $f(x)$ within these intervals to remove the discontinuities, and use the results above. The theorem would follow from a realization that the probability structure would be changed very little. The same device could be used to extend the results to cases where $f(x)$ is unbounded.

3. Application of results to tests of fit. First we prove the following lemma: If $F(x)$ and $G(x)$ are continuous distribution functions with density functions $f(x)$ and $g(x)$ respectively, then $f[F^{-1}(x)] \equiv cg[G^{-1}(x)]$ for some $c > 0$ if and only if $F(x) \equiv G(cx + b)$ for some constant b. To prove this, we note that the fact that $F[F^{-1}(x)] = x$ gives by differentiation that $f[F^{-1}(x)] \cdot (d/dx)F^{-1}(x) = 1$,

and $g[G^{-1}(x)]\,(d/dx)\,G^{-1}(x) \equiv 1$. Thus, if $f[F^{-1}(x)] \equiv cg[G^{-1}(x)]$, then $c\,(d/dx)\,F^{-1}(x) \equiv (d/dx)\,G^{-1}(x)$, so $cF^{-1}(x) \equiv G^{-1}(x) + B$, for some constant B. Letting $x = F(y)$, we get $cy = G^{-1}[F(y)] + B$, or $cy - B \equiv G^{-1}[F(y)]$, or $G(cy - B) \equiv F(y)$. Conversely, if $G(cx + b) \equiv F(x)$, then $cg(cx + b) \equiv f(x)$, while $cx + b \equiv G^{-1}[F(x)]$, so that $cg(G^{-1}[F(x)]) \equiv f(x)$, or setting $y = F(x)$, $cg(G^{-1}(y)) \equiv f[F^{-1}(y)]$, completing the proof of the lemma.

Now we examine the theorem of Sec. 2 for the special case $k = 1$, $g(u) = u$ (therefore $r = 1$), and $\rho(v) = (1/\beta)f[F^{-1}(v)]$, where β is a positive constant. Then $U(t) = 1 - e^{-t}$, and $S(t) = \int_{-\infty}^{\infty}[1 - e^{-\beta t}]f(x)\,dx = 1 - e^{-\beta t}$, and thus does not depend on $f(x)$. Suppose we are confronted with the following problem in hypothesis testing: $X_1, X_2, \cdots, X_n$ are known to be independent and identically distributed chance variables, with a continuous distribution, and the hypothesis is that the common distribution function is $F(cx + b)$ for some unknown constants c, $b (c > 0)$, where the form of $F(x)$ is known. Here c is a scale parameter, and b is a location parameter. We are going to examine the properties of the test which rejects the hypothesis when $\inf_{\beta>0} \sup_{t\geq 0} |R_n(t) - (1 - e^{-\beta t})|$ is "too large." We are going to show that this last expression converges stochastically to zero if and only if the hypothesis is true, so that the test is consistent. Also, when the hypothesis is true, the distribution of the expression is independent of the parameters c, b.

When the hypothesis is true, there is some $\alpha > 0$ such that $\sup_{t\geq 0} |R_n(t) - (1 - e^{-\alpha t})|$ converges stochastically to zero as n increases. This follows from the lemma at the beginning of this section, and implies that when the hypothesis is true, $\inf_{\beta>0} \sup_{t\geq 0} |R_n(t) - (1 - e^{-\beta t})|$ converges stochastically to zero as n increases. If the hypothesis is not true, then the true common distribution is $H(x)$, say, with density $h(x)$. Then, defining $S(t)$ as

$$\int_{-\infty}^{\infty}\left[1 - \exp\left\{\frac{-\beta t h(x)}{f[F^{-1}(H(x))]}\right\}\right] h(x)\,dx,$$

$\sup_{t\geq 0} |R_n(t) - S(t)|$ converges stochastically to zero as n increases. But $S(t)$ will equal $1 - e^{-\alpha t}$ for some positive α if and only if the hypothesis is true, and therefore when the hypothesis is not true, $\inf_{\beta>0} \sup_{t\geq 0} |R_n(t) - (1 - e^{-\beta t})|$ will not converge stochastically to zero. The fact that the distribution of $\inf_{\beta>0} \sup_{t\geq 0} |R_n(t) - (1 - e^{-\beta t})|$ is independent of the parameters c, b when the hypothesis is true follows immediately from the fact that if A, B are constants $(A > 0)$, and $\bar{R}_n(t)$ is the expression defined in terms of $AX_1 + B$, $AX_2 + B$, $\cdots$, $AX_n + B$ in exactly the same way as $R_n(t)$ was defined in terms of $X_1, X_2, \cdots, X_n$, then $\inf_{\beta>0} \sup_{t\geq 0} |\bar{R}_n(t) - (1 - e^{-\beta t})|$ is equal to

$$\inf_{\beta>0} \sup_{t\geq 0} |R_n(t) - (1 - e^{-\beta t})|.$$

REFERENCE

[1] LIONEL WEISS, "The stochastic convergence of a function of sample successive differences," *Ann. Math. Stat.*, Vol. 26 (1955), pp. 532–536.

THE ASYMPTOTIC POWER OF CERTAIN TESTS OF FIT BASED ON SAMPLE SPACINGS[1]

By Lionel Weiss

Cornell University

1. Introduction and summary. Suppose $X_1, X_2, \cdots, X_n$ are independent and identically distributed chance variables, each with density $f(x)$, where $\int_0^1 f(x)\,dx = 1$, $f(x)$ has a finite number of discontinuities, and there are two constants A, $B (0 < A < B < \infty)$ such that $A \leqq f(x) \leqq B$ for all x in $[0, 1]$.

Let Y_0 denote zero, Y_{n+1} denote unity, and let $Y_1 \leqq Y_2 \leqq \cdots \leqq Y_n$ be the ordered values of $X_1, X_2, \cdots, X_n$. Define T_i as $Y_i - Y_{i-1}$ for $i = 1, \cdots, n+1$. Let r be any positive number greater than unity, and let $V(n)$ denote $\sum_{i=1}^{n+1} T_i^r$. The following theorem was proved in [1].

Theorem A. *If $f(x) = 1$ for x in $[0, 1]$, then the distribution of*

$$\frac{n^{r-1/2}V(n) - \sqrt{n}\Gamma(r+1)}{\sqrt{\Gamma(2r+1) - (r^2+1)[\Gamma(r+1)]^2}}$$

approaches the standard normal distribution as n increases. In the present paper, we prove the following generalization of Theorem A:

Theorem 1: *The distribution of*

$$\frac{n^{r-1/2}V(n) - \sqrt{n}\Gamma(r+1)\int_0^1 f^{1-r}(x)\,dx}{\sqrt{[\Gamma(2r+1) - 2r\Gamma^2(r+1)]\int_0^1 f^{1-2r}(x)\,dx - \left[(r-1)\Gamma(r+1)\int_0^1 f^{1-r}(x)\,dx\right]^2}}$$

approaches the standard normal distribution as n increases.

Theorem 1 can be used to compute the asymptotic power of certain tests of fit based on $V(n)$.

2. Proof of Theorem 1 when $f(x)$ is a step function. First we prove Theorem 1 for the case when there are H subintervals $I_1, \cdots, I_H$, $I_1 = [0, z_1)$, $I_2 = [z_1, z_2), \cdots, I_H = [z_{H-1}, 1]$, so that on I_i, $f(x) = a_i$, where $0 < A \leqq a_i \leqq B$. Let N_i denote the number of the values $X_1, \cdots, X_n$ which fall in the interval I_i, and let ${}_iY_1 \leqq {}_iY_2 \leqq \cdots \leqq {}_iY_{N_i}$ be the ordered values of these values in I_i. Denote z_{i-1} by ${}_iY_0$, and z_i by ${}_iY_{N_i+1}$. z_0 is to denote zero, z_H denotes unity. Define ${}_iT_j$ as ${}_iY_j - {}_iY_{j-1}$, for $j = 1, \cdots, N_i + 1$. Define V_i as $\sum_{j=1}^{N_i+1} {}_iT_j^r$. From Theorem A quoted above and from an examination of the conditional distribution of ${}_iY_1, \cdots, {}_iY_{N_i}$ given N_i, it follows that the conditional distribution given N_i of

$$Q_i = \frac{\dfrac{N_i^{r-1/2}V_i}{(z_i - z_{i-1})^r} - \sqrt{N_i}\Gamma(r+1)}{\sqrt{\Gamma(2r+1) - (r^2+1)\Gamma^2(r+1)}}$$

Received July 23, 1956; revised July 26, 1956.

[1] Research under contract with the Office of Naval Research. It may be reproduced in whole or in part for any purpose of the United States Government.

approaches the standard normal distribution as N_i increases. Also, the conditional distribution of ${}_iY_1, \cdots, {}_iY_{N_i}$ given $N_1, \cdots, N_H$ depends only on N_i, while the joint distribution of $N_1, \cdots, N_H$ is multinomial with parameters $n, a_1(z_1 - z_0), \cdots, a_H(z_H - z_{H-1})$. From these facts, it follows that the joint distribution of

$$\left\{\sqrt{n}\left(\frac{N_1}{n} - a_1(z_1 - z_0)\right), \cdots, \sqrt{n}\left(\frac{N_{H-1}}{n} - a_{H-1}(z_{H-1} - z_{H-2})\right), Q_1, \cdots, Q_H\right\}$$

approaches the joint distribution of $\{S_1, \cdots, S_{H-1}, T_1, \cdots, T_H\}$ as n increases, where this last set of chance variables has a joint normal distribution with zero means and covariance matrix $\| a_{ij} \|$ $(i, j = 1, \cdots, 2H - 1)$, where $a_{ii} = 1$ if $i \geqq H$, $a_{ij} = 0$ if i and/or j is $\geqq H$, $a_{ii} = a_i(z_i - z_{i-1})[1 - a_i(z_i - z_{i-1})]$ if $i < H$, and $a_{ij} = -a_i a_j(z_i - z_{i-1})(z_j - z_{j-1})$ if i, j are both $< H$ and $i \neq j$.

Now $V(n)$ is equal to

$$\sum_{i=1}^{H} V_i - \sum_{1}^{H-1} {}_iT_{N_i+1}^r - \sum_{2}^{H} {}_iT_1^r + \sum_{1}^{H-1} [{}_iT_{N_i+1} + {}_{i+1}T_1]^r. \tag{2.1}$$

It can be verified easily from (2.1) and an examination of the distribution of ${}_iT_j$ that $n^{r-1/2}[V(n) - \sum_1^H V_i]$ converges stochastically to zero as n increases. Therefore if

$$\frac{n^{r-1/2}\left[\sum_1^H V_i\right] - \sqrt{n}\Gamma(r+1)\int_0^1 f^{1-r}(x)\,dx}{\sqrt{[\Gamma(2r+1) - 2r\Gamma^2(r+1)]\int_0^1 f^{1-2r}(x)\,dx - \left[(r-1)\Gamma(r+1)\int_0^1 f^{1-r}(x)\,dx\right]^2}} \tag{2.2}$$

has a limiting standard normal distribution as n increases, Theorem 1 is proved when $f(x)$ is a step function. Let us denote $\sqrt{n}[(N_i/n) - a_i(z_i - z_{i-1})]$ by W_i, and note that $W_1 + \cdots + W_H$ is identically equal to zero. The numerator of (2.2) can be written as

$$\begin{aligned}&\sqrt{\Gamma(2r+1) - (r^2+1)\Gamma^2(r+1)}\sum_1^H \frac{(z_i - z_{i-1})^r}{\left(\frac{N_i}{n}\right)^{r-1/2}} Q_i \\ &+ \sqrt{n}\Gamma(r+1)\sum_1^H \frac{(z_i - z_{i-1})}{\left[\frac{a_i N_i}{n}\right]^{r-1}}\left[[a_i(z_i - z_{i-1})]^{r-1} - \left[\frac{W_i}{\sqrt{n}} + a_i(z_i - z_{i-1})\right]^{r-1}\right]\end{aligned} \tag{2.3}$$

and remembering that N_i/n converges to $a_i(z_i - z_{i-1})$ with probability one as n increases, (2.3) has the same limiting distribution as

$$\begin{aligned}&\sqrt{\Gamma(2r+1) - (r^2+1)\Gamma^2(r+1)}\sum_1^H \frac{(z_i - z_{i-1})^{1/2}}{a_i^{r-1/2}} Q_i \\ &- (r-1)\Gamma(r+1)\sum_1^H \frac{W_i}{a_i^r}.\end{aligned} \tag{2.4}$$

But from the discussion above, it is easily verified that the distribution of (2.4) approaches a normal distribution with mean zero and variance equal to the square of the denominator of (2.2). This proves Theorem 1 when $f(x)$ is a step function.

3. Proof of Theorem 1 in the general case. The proof in the general case seems to require a great number of details, which we merely outline. In the first place, we may assume that $f(x)$ is continuous on $[0, 1]$, for if it has a finite number of discontinuities, we may handle each subinterval on which it is continuous separately, and then put them together as in Sec. 2. Then, defining λ_i as $|F(Y_i) - i/n|$, and remembering that $f(x) \geqq A > 0$, we find that $|Y_i - F^{-1}(i/n)| \leqq \lambda_i/A$. We have $F(Y_{i+1}) - F(Y_i) = f(\theta_i)[Y_{i+1} - Y_i]$, where $Y_i < \theta_i < Y_{i+1}$, or $F^{-1}(i/n) - (\lambda_i/A) < \theta_i < F^{-1}((i+1)/n) + (\lambda_{i+1}/A)$. Then we may write

$$F(Y_{i+1}) - F(Y_i) = f\left[F^{-1}\left(\frac{i}{n}\right)\right][Y_{i+1} - Y_i] + \gamma_i[Y_{i+1} - Y_i],$$

where $\gamma_i = f(\theta_i) - f[F^{-1}(i/n)]$. Due to the uniform continuity of $f(x)$, and the fact that $\max_i n^{1/2-\delta}\lambda_i$ converges stochastically to zero as n increases, we shall be able to ignore the term $\gamma_i[Y_{i+1} - Y_i]$ in certain respects. We denote $F(Y_{i+1}) - F(Y_i)$ by U_{i+1}, and $Y_{i+1} - Y_i$ by T_{i+1}. Then we may write

$$T_{i+1} = \frac{U_{i+1}}{f\left[F^{-1}\left(\frac{i}{n}\right)\right]} - \frac{\gamma_i T_{i+1}}{f\left[F^{-1}\left(\frac{i}{n}\right)\right]}. \tag{3.1}$$

We are going to examine the moments of the chance variable $W = \sum n^r T_i^r$, and it is clear from an examination of (3.1) that the leading terms of these moments will be the corresponding moments of, say,

$$\sum\left\{\frac{nU_i}{f\left[F^{-1}\left(\frac{i-1}{n}\right)\right]}\right\}^r = Q.$$

Let $V_1, \cdots, V_{n+1}$ be independent chance variables, each with density e^{-v} for $v > 0$. Then $E\{V_{i_1}^{a_1}V_{i_2}^{a_2} \cdots V_{i_k}^{a_k}\} = \Gamma(a_1 + 1)\Gamma(a_2 + 1) \cdots \Gamma(a_k + 1)$. Also, it is well known that

$$E\{(nU_{i_1})^{a_1}(nU_{i_2})^{a_2} \cdots (nU_{i_k})^{a_k}\} = \frac{(n^{a_1+\cdots+a_k})\Gamma(n+1)\Gamma(a_1+1) \cdots \Gamma(a_k+1)}{\Gamma(n + a_1 + \cdots + a_k + 1)},$$

and this last expression approaches $\Gamma(a_1 + 1) \cdots \Gamma(a_k + 1)$ as n increases. That is, with respect to their moments, the chance variables $nU_1, \cdots, nU_{n+1}$ act like the independent chance variables $V_1, \cdots, V_{n+1}$.

Defining the chance variable Q' as

$$\sum\left\{\frac{V_i}{f\left[F^{-1}\left(\frac{i-1}{n}\right)\right]}\right\}^r,$$

it is known that $E\{[(Q' - EQ') / \sigma_{Q'}]^k\}$ approaches μ_k, the kth moment of a standard normal chance variable, for any positive integral k. From the discussion above, one might expect the same to hold for $E\{[Q - EQ) / \sigma_Q]^k\}$, and a detailed examination shows that this is so. It is also so for $E\{[(W - EW) / \sigma_W]^k\}$, since the terms in this not given by the corresponding terms with W replaced by Q approach zero in the limit, due to the properties of γ_i defined above. This completes the proof.

4. The asymptotic power of certain tests of fit. To test the hypothesis that $f(x) = 1$ for $0 \leqq x \leqq 1$, the test that rejects when $V(n) \geqq C_n(\alpha)$ has been suggested, where $C_n(\alpha)$ is a constant depending on the sample size n and on the desired level of significance α. Denote $(1/\sqrt{2\pi})\int_v^\infty e^{-(t^2/2)}\,dt$ by $\phi(v)$, and let $k(\alpha)$ denote the value such that $\phi(k(\alpha)) = \alpha$. Then Theorem A shows that for large n, $C_n(\alpha)$ is approximately equal to

$$n^{-r+1/2}[\sqrt{n}\Gamma(r+1) + k(\alpha)\sqrt{\Gamma(2r+1) - (r^2+1)\Gamma^2(r+1)}],$$

while if the true common density is $f(x)$, then the large-sample power of the test is approximately equal to

$$\phi\left(\frac{n^{r-1/2}C_n(\alpha) - \sqrt{n}\Gamma(r+1)\int_0^1 f^{1-r}(x)\,dx}{\sqrt{[\Gamma(2r+1) - 2r\Gamma^2(r+1)]\int_0^1 f^{1-2r}(x)\,dx - \left[(r-1)\Gamma(r+1)\int_0^1 f^{1-r}(x)\,dx\right]^2}}\right).$$

REFERENCE

[1] D. A. DARLING, "On a class of problems related to the random division of an interval," *Ann. Math. Stat.*, Vol. 24 (1953), pp. 239–253.

THE DISTRIBUTION OF THE NUMBER OF LOCALLY MAXIMAL ELEMENTS IN A RANDOM SAMPLE

BY T. AUSTIN, R. FAGEN, T. LEHRER, AND W. PENNEY

Washington, D. C.

0. Summary. The distribution of the number of different locally maximal elements in a random sample is found, where the sampling is from a continuous population of real numbers. This distribution has application in certain nonparametric tests; the problem of finding the distribution may be regarded as identical with the enumeration of permutations according to the number of distinct locally maximal elements.

1. Introduction. An ordered sample of n real numbers is drawn at random from a population having a continuous distribution. For a given integer k, an element of the sample is called locally maximal if it is the largest of some k consecutive elements of the sample. The distribution of the number of different

Received September 4, 1956.

locally maximal elements in a random sample is discussed in the following; this distribution can be used as a basis for certain nonparametric tests in a manner analogous to other order statistics.

Although the problem arose in just such a context, it can, as will be indicated below, be treated as a purely combinatorial problem. The problem is then to enumerate the permutations on n objects according to the number of different locally maximal elements and so belongs to a class of problems similar to those studied by Riordan [1], Sprague [2], Sade [3], e.g., classifications of permutations according to various characteristics, such as rising sequences, falling sequences, readings [cf. Riordan [1]], etc.

2. Locally maximal elements. Let Ω be a population of real numbers with continuous distribution function, and let $0_n = (x_1, x_2, \cdots, x_n)$ be an ordered sample of size n drawn at random from Ω. For a given k let

$$y_i = \max(x_{i+1}, x_{i+2}, \cdots, x_{i+k}) \; i = 0, 1, 2, \cdots, n - k.$$

Let

$$z_j = \begin{cases} 1 & \text{if } y_i = x_j \text{ for at least one } i, \\ 0 & \text{otherwise} \end{cases} \qquad j = 1, 2, \cdots, n.$$

If $z_j = 1$, then x_j will be called a *k-maximal element*, or for brevity, a *maximal element*. Note that the probability of a tie, i.e., the event $x_l = x_m$ for $l \neq m$, is zero, and therefore there is no essential ambiguity in the definition of z_j. Clearly z_j is itself a random variable, being a function of a random sample. Thus the sequence $z_1, z_2, \cdots, z_n$ is a sequence of random variables (which are neither independent nor identically distributed) associated with the sample 0_n. Now let $S_n = \sum_{j=1}^{n} z_j$. The problem is to find the distribution of the random variable S_n, i.e., the set of numbers $\{p_s\}$,

$$p_s = P[S_n = s], \qquad s = 0, 1, 2, \cdots, n. \tag{1}$$

It is easily seen that the distribution of S_n is independent of the underlying distribution of Ω, and depends only on the order relationships among the numbers $x_1, \cdots, x_n$. It is convenient, therefore, to replace these numbers by the proper permutation of the integers $1, 2, \cdots, n$, i.e., that permutation having the same order relationships as $x_1, x_2, \cdots, x_n$. By symmetry, all permutations of $1, 2, \cdots, n$ have an equal probability of occurrence, so the distribution of S_n may be obtained by finding the number $f_k(n, i)$ of permutations of the first n integers which have exactly i different maximal elements.

3. Recurrence relationship and generating function for the numbers $f_k(n, i)$. A recurrence relation for the numbers $\{f_k(n, i)\}$ can be found in the following manner:

Consider all permutations of the first $n + 1$ integers in which the largest element is in the $(m + 1)$st position, i.e., those permutations of the form $a_1, a_2, \cdots, a_m, n + 1, a_{m+2}, \cdots, a_{n+1}$. Certain of these have exactly $i + 1$

different maximal elements. To enumerate such permutations, note that the element $(n+1)$ is necessarily maximal so that the permutations $(a_1, a_2, \cdots, a_m)$ and $(a_{m+2}, a_{m+3}, \cdots, a_{n+1})$ must between them contribute i different maximal elements. The m integers $a_1, \cdots, a_m$ which appear to the left of $(n+1)$ can be selected in $\binom{n}{m}$ ways; for any of these choices there are $f_k(m, r)$ permutations of $a_1, a_2, \cdots, a_m$ which have r maximal elements. Similarly, there are $f_k(n-m, i-r)$ permutations of the remaining integers $a_{m+2}, a_{m+3}, \cdots, a_{n+1}$ which have $i-r$ different maximal elements. Thus the total number of permutations of the first $n+1$ integers which have the largest element in the $(m+1)$st position and which have $i+1$ distinct maximal elements is

$$\sum_{r=0}^{i} f_k(m, r) f_k(n-m, i-r) \binom{n}{m}.$$

Summing on m, the total number, $f_k(n+1, i+1)$, of permutations of the first $n+1$ integers with $i+1$ different maximal numbers is given by

$$f_k(n+1, i+1) = \sum_{m=0}^{n} \sum_{r=0}^{i} f_k(m, r)\, f_k(n-m, i-r) \binom{n}{m}. \tag{2}$$

with the following boundary conventions:

$$\begin{cases} f_k(n, 0) = n!; & n < k, \\ f_k(n, 0) = 0\ ; & n \geqq k, \\ f_k(n, i) = 0\ ; & i > 0, n < k. \end{cases} \tag{3}$$

Note that, with these conventions, (2) holds for all n whenever $i > 0$ and for values of $n \geqq k-1$ when $i = 0$; (3) must be used to determine $f_k(n, i)$ for other values of n.

Using (2) and (3) the numbers $f_k(n, i)$ can be calculated recursively, and thus the desired distribution in (1) can be found for any fixed n since

$$p_s = P[S_n = s] = \frac{f_k(n, s)}{n!}. \tag{4}$$

Another way of generating the distribution in (4) arises from considering the generating function $v_k(x, y)$ of the numbers $f_k(n, i)\,/\,n!$. Let

$$v_k(x, y) = \sum_{\alpha=0}^{\infty} \sum_{\beta=0}^{\infty} \frac{f_k(\alpha, \beta)}{\alpha!} x^\alpha y^\beta. \tag{5}$$

From (5), it is obvious that

$$\left.\frac{\partial^n v_k}{\partial x^n}\right|_{x=0} = \sum_{\beta=0}^{\infty} f_k(n, \beta) y^\beta. \tag{6}$$

Equation (6) thus gives the generating function for the numbers $f_k(n, i)$ for fixed n, and hence the generating function for the entire distribution in (4) is just

$$\frac{1}{n!}\frac{\partial^n v_k}{\partial x^n}\bigg|_{x=0}.$$

Furthermore, it can be shown from (2), (3), and (5) that $v_k(x, y)$ satisfies the differential equation

$$\frac{\partial v}{\partial x} = yv^2 + (1 - y)(1 + 2x + 3x^2 + \cdots + (k - 1)x^{k-2}). \tag{7}$$

Also, note that from (2), $v_k(0, y) \equiv 1$, and so

$$\frac{\partial v_k}{\partial x}\bigg|_{x=0} = 1.$$

The generating function in (6) can therefore be found by repeated differentiation of (7). As a check, it can be shown by induction, using (3) and (6) for $y = 1$, that $\sum_{\beta=0}^{\infty} f_k(n, \beta) = n!$, as is of course necessary. Similarly one may show that for $n \geqq k$ the mean of S_n is given by

$$E(S_n) = \sum_{\beta=0}^{\infty} \frac{\beta f_k(n, \beta)}{n!} = \frac{1}{n!}\left[\frac{\partial}{\partial y}\left(\frac{\partial^n v_k}{\partial x^n}\right)\right]_{\substack{x=0 \\ y=1}} = \frac{2n - k + 1}{k + 1}.$$

These relations may also be derived, though less easily, by induction from (2) and (3).

4. Numerical examples. As an example, the first few values of $(\partial^n v_k / \partial x^n) \mid_{x=0}$ for $k = 3$, as found from (7), are given below:

$$\begin{cases} \dfrac{\partial v}{\partial x}\Big|_{x=0} = 1, \\ \dfrac{\partial^2 v}{\partial x^2}\Big|_{x=0} = 2, \\ \dfrac{\partial^3 v}{\partial x^3}\Big|_{x=0} = 6y, \\ \dfrac{\partial^4 v}{\partial x^4}\Big|_{x=0} = 12y + 12y^2, \\ \dfrac{\partial^5 v}{\partial x^5}\Big|_{x=0} = 24y + 72y^2 + 24y^3, \\ \dfrac{\partial^6 v}{\partial x^6}\Big|_{x=0} = 408y^2 + 264y^3 + 48y^4, \\ \dfrac{\partial^7 v}{\partial x^7}\Big|_{x=0} = 1008y^2 + 3120y^3 + 816y^4 + 96y^5, \\ \dfrac{\partial^8 v}{\partial x^8}\Big|_{x=0} = 2016y^2 + 18624y^3 + 17376y^4 + 2112y^5 + 192y^6. \end{cases} \tag{8}$$

The coefficients $f_k(n, \beta)$ of y^β in Eq. (8) can also be computed directly from (2).

REFERENCES

[1] JOHN RIORDAN, "Triangular permutation numbers," *Proc. Amer. Math. Soc.*, Vol. 2, No. 3, June 1951.

[2] R. S. SPRAGUE, "Uber ein Anordnungsproblem," *Math. Ann.*, Vol. 121 (1949).

[3] A. SADE, *Sur les chevauchements des permutations*, Marseille, 1949; and *Sur les suites hautes des permutations*, Marseille, 1949.

PERCOLATION PROCESSES: LOWER BOUNDS FOR THE CRITICAL PROBABILITY

BY J. M. HAMMERSLEY

United Kingdom Atomic Energy Research Establishment

1. Introduction. A percolation process is the spread of a fluid through a medium under the influence of a random mechanism associated with the medium. This contrasts with a diffusion process, where the random mechanism is associated with the fluid. Broadbent and Hammersley [1] gave examples illustrating the distinction.

Here we shall consider a *medium* consisting of an infinite set of *atoms* and *bonds*. A bond is a path between two atoms: it may be *undirected* (in which case it will allow passage from either atom to the other) or it may be *directed* (in which case it will allow passage from one atom to the other but not vice versa). Two atoms may be linked by several bonds, some directed and some undirected. Broadbent and Hammersley [1] dealt with *crystals*, i.e., media in which the atoms and bonds satisfied three postulates denoted by $P1$, $P2$, and $P3$. Here, however we shall dispense with $P1$ and a part of $P3$; and our surviving assumptions are:

$P2$. The number of bonds *from* (but not necessarily *to*) any atom is finite.

$P3$(a). Any finite subset of atoms contains an atom *from* which a bond leads to some atom not in the subset.

With this medium we associate the following *random mechanism*: each bond has an independent probability p of being *undammed* and $q = 1 - p$ of being *dammed*. *Fluid*, supplied to the medium at a set of *source atoms*, spreads along undammed bonds only (and in the permitted direction only for undammed directed bonds) and thereby *wets* the atoms it reaches. Associated with each atom A, there is a *critical probability* $p_d(A)$, defined as the supremum of all values of p such that, when A is the only source atom, A wets only finitely many atoms with probability one. We seek lower bounds for p_d.

An *n-stepped walk* is an ordered connected path along n bonds, each step being in a permitted direction along its bond and starting from the atom reached by the previous step. Walks (as opposed to fluid) may traverse dammed bonds: a walk is dammed or undammed according as it traverses at least one or no

Received October 29, 1956.

dammed bond. A walk is *self-avoiding* if it visits no atom more than once. The number of n-stepped self-avoiding walks starting from the atom A is denoted by $f_A(n)$. The *connective bound* of the medium is defined to be

$$\lambda = \sup_A \limsup_{n\to\infty} n^{-1} \log f_A(n). \tag{1}$$

Hammersley [3] showed that, under the fuller assumptions required by a crystal, there existed a *connective constant*

$$\kappa = \lim_{n\to\infty} n^{-1} \log f_A(n), \tag{2}$$

independent of A. Clearly $\lambda = \kappa$ when the latter exists.

The *principal n-neighbourhood* of an atom A, written $\mathbf{N}^n(A)$, is the atom A together with all atoms accessible from A by walks of n or fewer steps. The *principal n-boundary* of A is $\mathbf{B}^n(A) = \mathbf{N}^n(A) - \mathbf{N}^{n-1}(A)$, where $\mathbf{N}^0(A)$ is A alone. *A walk belongs to a set of atoms* $\mathbf{S}$ if every step of the walk *starts* from some atom of $\mathbf{S}$: notice that the final step may *terminate* at some atom not in $\mathbf{S}$.

2. Statement of results. Let $E_n(A, p)$ denote the expected number of atoms of $\mathbf{B}^n(A)$ which can be reached from A by at least one undammed walk belonging to $\mathbf{N}^{n-1}(A)$. Define $F_n = F_n(p) = \sup_A E_n(A, p)$.

THEOREM 1. $F_n(p_d + 0) \geqq 1$.

It is an easy matter to show from $P2$ and $P3(a)$ that $\mathbf{N}^n(A)$ has only finitely many atoms, and that every atom of $\mathbf{B}^n(A)$ can be reached from A by at least one (perhaps dammed) walk belonging to $\mathbf{N}^{n-1}(A)$, and that $\mathbf{B}^n(A)$ contains at least one atom. Therefore $E_n(A, p)$ is an increasing function of p, and $F_n(p)$ is a nondecreasing function of p, and $0 = E_n(A, 0) = F_n(0)$ and $1 \leqq E_n(A, 1) \leqq F_n(1)$. Thus Theorem 1 provides a lower bound for p_d, because p_d exceeds any solution of $F_n(p) < 1$.

Let $P_n = P_n(A)$ be the probability that the single source atom A wets at least one atom of $\mathbf{B}^n(A)$.

THEOREM 2. *If* $F_n < 1$ *for some particular* n, *then* $P_N \leqq F_n^{[N/n]}$ *for all* N, *where* $[N/n]$ *denotes the integer part of* N/n.

Theorem 2 is a rather more precise form of Theorem 1 and will be required elsewhere [5]. Here we shall deduce Theorem 1 from Theorem 2.

THEOREM 3. $p_d \geqq e^{-\lambda}$.

Theorem 3 is a straightforward generalization of a previous result ([1], Theorem 7).

I have not yet succeeded in proving or disproving

CONJECTURE 1. *For each fixed* p, F_n *is a subexponential function of* n, *that is to say,* $F_{l+m} \leqq F_l F_m$.

Example 1 below shows that Theorem 1 is sometimes stronger than Theorem 3. Theoretically, Theorem 1 is never weaker than Theorem 3. However, when the medium is a crystal it is not hard to estimate κ by Monte Carlo methods, and Theorem 3 may prove more useful than Theorem 1 in cases in which F_n is hard to calculate for large n. Similarly, it may occasionally happen that Theorem

1, although weakened theoretically thereby, may yet be strengthened practically by redefining E_n as the expected number of wet atoms in $\mathbf{B}^n(A)$.

EXAMPLE 1. Let the atoms of the medium lie in the Euclidean plane at the points (x, y), where $x, y = 0, 1, 2, \cdots$. Suppose that from each (x, y) there is a directed bond to $(x + 1, y)$ and a directed bond to $(x, y + 1)$. Then $f_4(n) = 2^n$, $\lambda = \log 2$, and Theorem 3 gives $p_d \geqq \frac{1}{2}$. However, $F_2(p) = 4p^2 - p^4$; so that Theorem 1 gives $p_d > \sqrt{\frac{1}{2}} - \sqrt{\frac{1}{2}} = 0.518 \cdots$. F_3 gives a slightly sharper result, but is much more tedious to calculate. We may also notice that $F_1(p) = 2p$; so that $F_2 < F_1^2$, and inequality is certainly required in Conjecture 1.

EXAMPLE 2. Consider the familiar branching process in which each individual has 0, 1, or 2 descendants with independent probabilities q^2, $2pq$, p^2. This, like all branching processes, is a special form of percolation process (see [1] for further details of this question). We consider the atoms of the medium to be the actual or potential individuals of the branching process. Each atom has just two directed bonds from it, and the fluid is life transforming a potential into an actual individual. We have $F_n = (2p)^n$ and $p_d = \frac{1}{2}$, agreeing with well-known results. Also $\lambda = \log 2$. Thus Theorems 1 and 3 are equally strong and best possible. Also Conjecture 1 holds with equality. This and other branching processes suggest

PROBLEM 1. *Under what conditions is* $\liminf_{n\to\infty} F_n(p_d) = 1$ *valid?*

We cannot in general use E in the role of F nor omit the $+0$ in Theorem 1, though perhaps the exceptional cases are rare. Example 3 provides one such exception.

EXAMPLE 3. Suppose that the atoms $A_1, A_2, \cdots$ are connected by $2j$ directed bonds from A_j to A_{j+1}. Then

$$E_n(A_m, p) = \prod_{j=m}^{m+n-1} (1 - q^{2j}) < 1$$

for $p < 1$. However $p_d = 0$, because $E_\infty(A_m, p)$ is also the probability that A_m wets infinitely many atoms and the infinite product converges for $q < 1$. As a matter of passing interest, $E_\infty(A_1, p) = (\vartheta_1'/2q^{1/4})^{1/8}$, where q plays the usual theta-function role of Jacobi's nome [6], p. 473; see also [2], Sec. 21.7: indeed, $2j$ rather than j bonds from A_j preserved the nomenclature.

3. Proof of theorems. Let A_1 be a fixed atom and n a fixed positive integer. In studying the spread of fluid from the single source atom A_1, we shall suppose that the spreading occurs in consecutive recursively defined stages. Immediately before the jth stage takes place, we shall know two sets of atoms, denoted by $\mathbf{W}(j)$ and $\mathbf{S}(j)$ respectively. Here $\mathbf{W}(j)$ is the set of atoms already wet up to but excluding the jth stage, and $\mathbf{S}(j)$ is the set of atoms about to serve as sources in the jth stage. The process starts from $\mathbf{W}(1) = \mathbf{S}(1) = A_1$. If $\mathbf{S}(j)$ is empty, then $\mathbf{S}(j + 1)$ is empty, and $\mathbf{W}(j + 1) = \mathbf{W}(j)$. If $\mathbf{S}(j)$ is not empty, let A be any atom of $\mathbf{S}(j)$ and proceed as follows. Define $\mathbf{X}(A)$ to be the set of all atoms in $\mathbf{N}^n(A) - \mathbf{W}(j)$, which can be reached from A by at least one undammed

self-avoiding walk belonging to $[\mathbf{N}^{n-1}(A) - \mathbf{W}(j)] + A$. Define $\mathbf{Y}(A)$ as the intersection of $\mathbf{X}(A)$ and $\mathbf{B}^n(A)$. Lastly define

$$\mathbf{S}(j+1) = \sum_A \mathbf{Y}(A); \qquad \mathbf{W}(j+1) = \mathbf{W}(j) + \sum_A \mathbf{X}(A), \tag{3}$$

where $\sum_A$ denotes summation over all atoms A belonging to $\mathbf{S}(j)$. We shall also require sets of atoms $\mathbf{T}(1)$, $\mathbf{T}(2)$, $\cdots$ defined recursively by

$$\mathbf{T}(1) = A_1; \qquad \mathbf{T}(j+1) = \sum_{A \in \mathbf{T}(j)} \mathbf{B}^n(A), \qquad j = 1, 2, \cdots. \tag{4}$$

Let A be an atom in $\mathbf{S}(j)$, supposed not empty, and let B be an atom of $\mathbf{T}(j+1)$. We write $A \to B$ to denote the existence of at least one undammed self-avoiding walk from A to B belonging to $[\mathbf{N}^{n-1}(A) - \mathbf{W}(j)] + A$. By the foregoing definitions, $A \to B$ implies that $B \in \mathbf{S}(j+1)$; and conversely there exists an atom $A_j \in \mathbf{S}(j)$ if and only if we can find atoms $A_1, A_2, \cdots, A_{j-1}$ belonging to $\mathbf{S}(1), \mathbf{S}(2), \cdots, \mathbf{S}(j-1)$ such that $A_1 \to A_2 \to \cdots \to A_{j-1} \to A_j$. Notice also that the sets $\mathbf{S}(1), \mathbf{S}(2), \cdots$ are mutually disjoint; and that $\mathbf{S}(1), \mathbf{S}(2), \cdots, \mathbf{S}(j)$ are all subsets of $\mathbf{W}(j)$. Finally $\mathbf{S}(j)$ is a subset of $\mathbf{T}(j)$. It may also be a subset of $\mathbf{T}(k)$ for $k \neq j$; but this will not affect our argument.

If A is an atom of the nonempty set $\mathbf{S}(j)$, we define its score $\theta(A) = \sum_B \theta(B)$, where $\sum_B$ denotes summation over all atoms $B \in \mathbf{S}(j-1)$ such that $B \to A$. We begin this recursive definition from $\theta(A_1) = 1$ when $j = 1$. To the set $\mathbf{T}(j)$ we attach the score

$$\phi_j = \begin{cases} \sum_{A \in \mathbf{S}(j)} \theta(A) & \text{if } \mathbf{S}(j) \text{ is not empty,} \\ 0 & \text{if } \mathbf{S}(j) \text{ is empty.} \end{cases} \tag{5}$$

Suppose that $\mathbf{W}(j-1)$ is given, and that $\mathbf{S}(j-1)$ is not empty. This means that $\mathbf{S}(1), \mathbf{S}(2), \cdots, \mathbf{S}(j-1)$ are all given and not empty. Hence $\theta(B)$ is given for each $B \in \mathbf{S}(j-1)$. Consider the conditional expectation of ϕ_j given $\mathbf{W}(j-1)$ with nonempty $\mathbf{S}(j-1)$. We have

$$\begin{aligned} E[\phi_j \mid \mathbf{W}(j-1), \quad & \mathbf{S}(j-1) \neq 0] \\ &= \sum_{A \in \mathbf{T}(j), B \in \mathbf{S}(j-1)} \theta(B) \text{ Prob } [B \to A \mid \mathbf{W}(j-1)] \\ &= \sum_{B \in \mathbf{S}(j-1)} \theta(B) \sum_{A \in \mathbf{T}(j)} \text{Prob } [B \to A \mid \mathbf{W}(j-1)]. \end{aligned} \tag{6}$$

Since $B \to A$ involves the existence of at least one undammed self-avoiding walk from B to A belonging to $[\mathbf{N}^{n-1}(B) - \mathbf{W}(j-1)] + B$, this event depends only upon the condition of bonds whose condition does not affect $\mathbf{W}(j-1)$. Hence,

$$\text{Prob } [B \to A \mid \mathbf{W}(j \leqq 1)] \leqq \text{Prob } [B \sim A], \tag{7}$$

where $B \sim A$, in the unconditional probability on the right of (7), means that there is at least one undammed walk from B to A belonging to $\mathbf{N}^{n-1}(B)$. Then, by definition of $E_n(B, p)$ and $F_n(p)$, we have

$$\sum_{A \in \mathbf{T}(j)} \text{Prob } [B \sim A] = E_n(B, p) \leqq F_n(p). \tag{8}$$

Combination of (6), (7), and (8) yields

$$E[\phi_j \mid \mathbf{W}(j-1), \quad \mathbf{S}(j-1) \neq \mathbf{0}] \leqq F_n(p) \sum_{B \in \mathbf{S}(j-1)} \theta(B). \tag{9}$$

In (9), we can remove the condition $S(j-1) \neq 0$, provided we interpret the right-hand empty sum as zero when $\mathbf{S}(j-1) = \mathbf{0}$. Hence,

$$\begin{aligned} E[\phi_j] &= \sum E[\phi_j \mid \mathbf{W}(j-1)] \text{ Prob } [\mathbf{W}(j-1)] \\ &\leqq F_n(p) \sum_{B \in \mathbf{S}(j-1)} \sum \theta(B) \text{ Prob } [\mathbf{W}(j-1)] = F_n(p) E[\phi_{j-1}]. \end{aligned} \tag{10}$$

Since *every* walk from A_1 to $\mathbf{B}^N(A_1)$ contains at least N steps, A_1 cannot wet any atom of $\mathbf{B}^N(A_1)$ unless none of $\mathbf{S}(1)$, $\mathbf{S}(2)$, $\cdots$, $\mathbf{S}(\nu+1)$ are emptv. where $\nu = [N/\nu]$. If $\mathbf{S}(\nu+1)$ is not empty, $\phi_{\nu+1} \geqq 1$. Thus, by (10),

$$P_N \leqq \text{Prob } [\phi_{\nu+1} \geqq 1] \leqq E[\phi_{\nu+1}] \leqq F_n^{\nu} = F_n^{[N/n]}, \tag{11}$$

which is Theorem 2. The relation (11) is true but useless if $F_n \geqq 1$.

If

$$F_n = F_n(p) < 1, \tag{12}$$

then

$$\lim_{N \to \infty} P_n = 0, \tag{13}$$

by (11). Since $\mathbf{N}^N(A_1)$ contains only finitely many atoms, (13) implies that A_1 wets infinitely many atoms with probability zero. Therefore, by definition of $p_d = p_d(A_1)$,

$$p \leqq p_d\,. \tag{14}$$

Since (14) is a consequence of (12), we deduce Theorem 1.

Theorem 3 is easy; for A_1 does not wet $\mathbf{B}^N(A_1)$ unless there is at least one N-stepped self-avoiding walk from A_1. The probability of this event is less than or equal to the expected number of such walks, namely $p^N f_A(N)$, because all bonds of a self-avoiding walk are distinct. If $p < e^{-\lambda}$, $\lim_{n \to \infty} p^N f_A(N) = 0$ by (1), and Theorem 3 follows.

To see that Theorem 1 is always as strong as Theorem 3, notice that, given $\epsilon > 0$, there exists n such that $f_A(m) \leqq e^{(\lambda+\epsilon)m}$ for $m \geqq n$; and hence $E_n(A, p) \leqq \sum_{m \geqq n} (pe^{\lambda+\epsilon})^m$. The right-hand side does not depend on A, so that we may write $F_n(p)$ for $E_n(A, p)$. Then, if $p < e^{-(\lambda+\epsilon)}$, $F_n \to 0$ as $n \to \infty$; and the result follows because ϵ is arbitrary.

REFERENCES

[1] S. R. Broadbent and J. M. Hammersley, "Percolation processes. I. Crystals and mazes," (To appear, *Proc. Camb. Phil. Soc.*)

[2] A. Fletcher, J. C. P. Miller and L. Rosenhead, *An index of mathematical tables*, Scientific Computing Service Ltd., London, 1946.

[3] J. M. HAMMERSLEY, "Percolation processes. II. The connective constant," (To appear, *Proc. Camb. Phil. Soc.*)
[4] J. M. HAMMERSLEY, "Percolation processes, III. Gravity crystals," unpublished.
[5] J. M. HAMMERSLEY, "Percolation processes. V. Upper bounds for the critical probability," unpublished.
[6] E. T. WHITTAKER AND G. N. WATSON, *Modern Analysis*, 4th ed., Cambridge University Press, 1950.

NON-PARAMETRIC UP-AND-DOWN EXPERIMENTATION[1]

BY CYRUS DERMAN

Columbia University

1. Introduction. Let $Y(x)$ be a random variable such that $P(Y(x) = 1) = F(x)$ and $P(Y(x) = 0) = 1 - F(x)$ where $F(x)$ is a distribution function. It is sometimes of interest, as in sensitivity experiments, to estimate a given quantile of $F(x)$ with observations distributed like $Y(x)$ where the choice of x is under control. A procedure for estimating the median was suggested by Dixon and Mood [2]. The validity of their procedure depends on the assumption that $F(x)$ is normal. Robbins and Monro [6] suggested a general scheme which can be used for estimating any quantile and which imposes no parametric assumptions on $F(x)$. Their method does assume, however, that the range of possible experimental values of x is the real line. In practice, this will not be the case. Limitations on the precision of measuring instruments, or natural limitations such as when x is obtained by a counting procedure, will usually restrict the experimental range of x to a set of numbers of the form

$$a + hn (-\infty < a < \infty, h > 0, n = 0, \pm 1, \cdots).$$

In this note we suggest a non-parametric procedure for estimating any quantile of $F(x)$ on the basis of quantal response data when, experimentally, x is restricted to the form $a + hn$.

For convenience we assume $a = 0$, $h = 1$. Suppose we wish to estimate that value of $x = \theta$ such that $F(\theta - 0) \leqq \alpha \leqq F(\theta)$, $\frac{1}{2} \leqq \alpha < 1$. If $0 < \alpha \leqq \frac{1}{2}$ or $a \neq 0$ or $h \neq 1$ the necessary modifications will be apparent. The experimental procedure is as follows: choose x_1 arbitrarily. Recursively, let

$$\begin{aligned} x_n &= x_{n-1} - 1, && \text{with probability } \frac{1}{2\alpha} \text{ if } y_{n-1} = 1, \\ &= x_{n-1} + 1, && \text{with probability } 1 - \frac{1}{2\alpha} \text{ if } y_{n-1} = 1, \\ &= x_{n-1} + 1, && \text{with probability 1 if } y_{n-1} = 0. \end{aligned} \tag{1}$$

Received May 28, 1956; Revised January 29, 1957.
[1] Research supported by the United States Air Force through the Office of Scientific Research of the Air Research and Development Command.

where y_k denotes the zero-or-one response at x_k. The estimate θ_n of θ based on n observations is the *most frequent* value of x, if unique, or the *arithmetric average of the most frequent levels*, if not unique.

We shall prove the following

THEOREM. *If $F(x)$ is strictly increasing for $\theta - 1 \leqq x \leqq \theta + 1$, then*

$$P(\max(|\limsup_{n\to\infty} \theta_n - \theta|, |\liminf_{n\to\infty} \theta_n - \theta|) < 1) = 1.$$

2. Two lemmas.

Let $\{X_n\}(n = 0, 1, \cdots)$ be an irreducible Markov chain with recurrent non-null states and stationary transition probabilities $\{p_{ij}\}$ (see Feller [3] for definitions of terms) such that

$$p_{i,i+1} + p_{i,i-1} = 1 \qquad (i = 0, \pm 1, \cdots). \tag{2}$$

Let v_i $(i = 0, \pm 1, \cdots)$ be the unique solution of the equations

$$\begin{cases} \sum_{i=-\infty}^{\infty} v_i p_{ij} = v_j & (j = 0, \pm 1, \cdots), \\ v_i > 0, & \text{for all } i, \\ \sum_{i=-\infty}^{\infty} v_i = 1. \end{cases} \tag{3}$$

Since $\{X_n\}$ is irreducible and the states are recurrent non-null, the system (3) has such a unique solution. The v_i's play the role of stationary absolute probabilities; i.e., if $P(X_0 = i) = v_j$, then $P(X_n = i) = v_i$ for every n.

LEMMA 1. *If for some $i = b$, $p_{b,b+1} \leqq p_{b,b-1}$, $p_{b,b+1} > p_{b+1,b+2}$ and $p_{i,i+1}$ is non-increasing in i for $i \geqq b + 1$, then $v_b > b_{b+1}$ and v_i is non-increasing in i for $i \geqq b + 1$. Similarly, if for some $i = c$, $p_{c,c-1} \leqq p_{c,c+1}$, $p_{c,c-1} > p_{c-1,c-2}$, and $p_{i,i+1}$ is non-decreasing in i for $i \leqq c - 1$, then $v_c > v_{c-1}$ and v_i is non-decreasing in i for $i \leqq c - 1$.*

Proof. Let $\pi_{ij} = P(X_n = j \text{ for some } n \geqq 1, X_r \neq i \text{ or } j \text{ for } r < n \mid X_0 = i)$. From a result of Harris [5] we know that

$$\frac{v_{i+1}}{v_i} = \frac{\pi_{i,i+1}}{\pi_{i+1,i}}. \tag{4}$$

It is clear however that $\pi_{i,i+1} = p_{i,i+1}$ and $\pi_{i+1,i} = p_{i+1,i}$. Hence, from (4) and by the hypothesis

$$\frac{v_{b+1}}{v_b} = \frac{p_{b,b+1}}{p_{b+1,b}} = \frac{p_{b,b+1}}{1 - p_{b+1,b+2}} < \frac{p_{b,b+1}}{1 - p_{b,b+1}} \leqq 1$$

and thus $v_{b+1} < v_b$. The remainder of the proof follows in the same manner.

Let $N_n(i)$ denote the number of r such that $X_r = i$ for $r \leqq n$. For the truth of the following lemma we need not impose the condition (2).

LEMMA 2. *Let B be the set of states such that $v_{i'} = \max_i \{v_i\}$ for $i' \varepsilon B$. Then for every $i' \varepsilon B$.*

$$P\left(\lim_{n\to\infty} \frac{N_n(i')}{n} = v_{i'} > \lim_{n\to\infty} \max_{i \notin B} \left\{\frac{N_n(i)}{n}\right\}\right) = 1.$$

Proof. Since $\sum_{i=-\infty}^{\infty} v_i = 1$, there exists a finite set A of states with $B \subset A$ such that $\sum_{i \in A} v_i < v_{i'}$. From the strong law of large numbers for Markov chains [1], it follows that $P(\lim_{n\to\infty} (N_n(i)/n = v_i) = 1$ for every i and more generally $P(\lim_{n\to\infty} \sum_{i \in A} (N_n(i)/n = \sum_{i \in A} v_i) = 1$. Let ϵ be any number such that $0 < \epsilon < v_{i'} - \max(\max_{i \in A-B} \{v_i\}, \sum_{i \in A} v_i)$ and let E_N denote the event that $(N_n(i') / n > v_{i'} - \epsilon$ for all $n > N$. By the previous remark and since $\{E_N\}$ is a monotone sequence, $\lim_{N\to\infty} P(E_N) = P(\lim_{N\to\infty} E_N) = 1$. Therefore there exists an N_1 such that $P(N_n(i') / n > v_{i'} - \epsilon$ for all $n > N_1) > 1 - \epsilon/3$. Similarly, since A is finite, there exists an N_2 such that $P(\max_{i \in A-B} \{N_n(i) / n\} < v_{i'} - \epsilon$ for all $n > N_2) > 1 - \epsilon/3$ and an N_3 such that $P(\sum_{i \in A} N_n(i) / n < v_{i'} - \epsilon$ for all $n > N_3) > 1 - \epsilon/3$. Let $N_0 = \max(N_1, N_2, N_3)$. Then it follows that

$$P(N_n(i') / n > v_{i'} - \epsilon$$

$$> \max(\max_{i \in A-B} \{N_n(i) / n\}, \sum_{i \in A} N_n(i) / n) \text{ for all } n > N_0) > 1 - \epsilon.$$

Since $\epsilon > 0$ is arbitrary, we have

$$P(\lim_{n\to\infty} N_n(i') / n = v_{i'} > \limsup_{n\to\infty} \max_{i \in B} \{N_n(i) / n\}) = 1.$$

The last assertion implies that $\lim_{n\to\infty} \max_i \{N_n(i) / n\}$ exists. By a similar argument applied to the finite set B_1 of states which have the second largest v_i's it follows that $\limsup_{n\to\infty} \max_{i \in B} \{N_n(i) / n\}$ can be replaced by $\lim_{n\to\infty} \max_{i \in B} \{N_n(i)/n\}$. The lemma is proved.

3. Application of lemmas.

Let $\{X_n\}$ be the Markov chain defined by (1); i.e. let $X_n = i$ if $x_n = i$. The transition probabilities are of the form

$$p_{i,i+1} = 1 - \frac{F(i)}{2\alpha},$$

$$p_{i,i-1} = \frac{F(i)}{2\alpha}.$$

The chain is clearly irreducible and the states can be easily shown to be recurrent non-null using a theorem of Harris [5] or a modified version of a theorem of Foster [4]. The numbers $[\theta] + 1$ and $[\theta]$, where $[\theta]$ denotes the largest integer less than or equal to θ, can be taken as b and c of Lemma 1. From Lemma 1 and the condition of strict monotonicity of $F(x)$ for $\theta - 1 \leqq x \leqq \theta + 1$, it is clear that $[\theta]$ or $[\theta] + 1$ or both but no other states belong to B of Lemma 2. Thus, according to Lemma 2, the most frequent state, for n large enough, will be $[\theta] + 1$, $[\theta]$ or both with probability 1. In any case, the difference between θ and $[\theta] + 1$ or $[\theta]$ or the arithmetic average of the two is less than 1. The theorem is therefore proved.

REFERENCES

[1] K. L. Chung, "Contributions to the theory of Markov chains," *Trans. Amer. Math. Soc.*, Vol. 76, No. 3 (1954), pp. 397–419.

[2] W. J. Dixon and A. M. Mood, "A method for obtaining and analyzing sensitivity data," *J. Amer. Stat. Assn.*, Vol. 43 (1948), pp. 109–126.
[3] W. Feller, *Probability Theory and Its Applications*, John Wiley & Sons, Inc., New York, 1950.
[4] F. G. Foster, "Markoff chains with an enumerable number of states and a class of cascade processes," *Proc. Cambridge Philos. Soc.*, Vol. 47 (1951), pp. 77–85.
[5] T. E. Harris, "First passage and recurrence distributions," *Trans. Amer. Math. Soc.*, Vol. 73, No. 3 (1952), pp. 471–486.
[6] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Stat.*, Vol. 22 (1951), pp. 400–407.

APPROXIMATE MOMENTS FOR THE SERIAL CORRELATION COEFFICIENT

By John S. White[1]

Ball Brothers Co.

1. Introduction and summary. The first order Gaussian auto-regressive process (x_t) may be defined by the stochastic difference equation

$$x_t = \rho x_{t-1} + u_t , \tag{1}$$

where the u's are NID(0, 1) and ρ is an unknown parameter. The choice of a statistic as an estimator for ρ depends on the initial conditions imposed on the difference equation (1). The so-called "circular" model is obtained by considering a sample of size N and then assuming that $x_{N+1} = x_1$. An appropriate estimator for ρ in this case is the circular serial correlation coefficient

$$r = \frac{\sum_{t=1}^{N} x_t x_{t+1}}{\sum_{t=1}^{N} x_t^2} \qquad (x_{N+1} = x_1). \tag{2}$$

Leipnik [1] has derived an approximate density function

$$f(t) = \frac{\Gamma\left(\frac{N+2}{2}\right)}{\Gamma\left(\frac{N+1}{2}\right)\Gamma\left(\frac{1}{2}\right)} (1 - 2t\rho + \rho^2)^{-N/2}(1 - t^2)^{(N-1)/2} \tag{3}$$

for the estimator r. Leipnik also evaluated the first two moments of this distribution. In this paper a formula is obtained which gives $E(r^k)$ as a polynomial of degree k in ρ.

2. The general formula for $E(r^k)$. To calculate the moments of r we must evaluate the integral

$$E(r^k) = \int_{-1}^{1} t^k f(t)\, dt. \tag{4}$$

Received August 13, 1956; revised February 24, 1957.
[1] Now with Aero Division, Minneapolis Honeywell Regulator Company, Minneapolis.

The direct integration of this function is not obvious; however, it can be evaluated quite easily by means of the Gegenbauer polynomials.

Gegenbauer's function $C_j^n(t)$ for integral values of j is defined to be the coefficient of ρ^j in the expansion of $(1 - 2t\rho + \rho^2)^{-n}$ in powers of ρ (for this and the following results concerning the Gegenbauer functions see Magnus and Oberhettinger [2] pp. 77 and 78).

$$(5) \qquad (1 - 2t\rho + \rho^2)^{-n} = \sum_{j=0}^{\infty} C_j^n(t)\rho^j.$$

The Gegenbauer polynomials are orthogonal over the interval $(-1, 1)$ with weight function $(1 - t^2)^{n-1/2}$ and have the general properties of the classical orthogonal polynomials.

One special result which we shall apply is the following. Let $g(t)$ be a continuous function with j continuous derivatives; then

$$(6) \qquad \int_{-1}^{1} g(t)(1 - t^2)^{n-1/2}C_j^n(t)\,dt = K(n, j)\int_{-1}^{1} (1 - t^2)^{j+n-1/2}\frac{d^j g(t)}{dt^j}\,dt,$$

where

$$K(n, j) = \frac{\Gamma(2n + j)\Gamma(n + \frac{1}{2})}{\Gamma(2n)\Gamma(n + j + \frac{1}{2})\Gamma(j + 1)2^j}.$$

This result may be verified by applying the "Rodrigues Formula" for $C_j^n(t)$ (see [2], p. 78, line 2) to the left-hand side of (6) and then integrating by parts j times.

Expanding the denominator of (4) in a series (5) we have

$$(7) \qquad E(r^k) = \frac{\Gamma\left(\frac{N+2}{2}\right)}{\Gamma\left(\frac{N+1}{2}\right)\Gamma\left(\frac{1}{2}\right)}\int_{-1}^{1} t^k(1 - t^2)^{(N-1)/2}\left[\sum_{j=0}^{\infty} C_j^{N/2}(t)\rho^j\right]dt.$$

Since, for this problem, $|\rho| < 1$ and $|t| \leqq 1$, (5) may be written as

$$(8) \qquad (1 - 2\rho\cos\theta + \rho^2)^{-n} = (1 - \rho e^{-i\theta})^{-n}(1 - \rho e^{i\theta})^{-n}$$

$$(8a) \qquad = \sum_{j=0}^{\infty} C_j^n(\cos\theta)\rho^j.$$

Expanding the right-hand side of (8) in powers of h as the product of two binomial series and comparing coefficients of h^j with those in (8a) we have

$$(9) \qquad |C_j^n(\cos\theta)| \leqq \binom{-2n}{j}.$$

Hence, by the Weierstrass M-test, the series $\sum C_j^n(\cos\theta)\rho^j = \sum C_j^n(t)\rho^j$ converges uniformly in t.

Since the series converges uniformly we may invert the order of integration and summation in (7). Applying (6) to (7) with $g(t) = t^k$ and $n = N/2$, we have

$$E(r^k) = \sum_{j=0}^{k} K(N/2, j) \frac{\Gamma\left(\frac{N+2}{2}\right)}{\Gamma\left(\frac{N+1}{2}\right)\Gamma\left(\frac{1}{2}\right)} \cdot \int_{-1}^{1} k(k-1)\cdots(k-j+1)t^{k-j}(1-t^2)^{j+(N-1)/2}\,dt. \tag{10}$$

We note that

$$\int_{-1}^{1} t^p(1-t^2)^q\,dt = 0 \qquad \text{for } p \text{ odd},$$

$$= \frac{\Gamma\left(\frac{p+1}{2}\right)\Gamma(q+1)}{\Gamma\left(\frac{p+1}{2}+q+1\right)} \qquad \text{for } p \text{ even}.$$

If we now let $2i = k - j$ ($i = 0, 1, \cdots, [k/2]$), (10) becomes

$$E(r^k) = \sum_{i=0}^{[k/2]} \frac{\Gamma(N+k-2i)\Gamma(k+1)\Gamma(i+\frac{1}{2})\Gamma([N+2]/2)\rho^{k-2i}}{\Gamma(N)\Gamma(k-2i+1)\Gamma(2i+1)\Gamma(\frac{1}{2})\Gamma\left(\frac{N+2}{2}+[k-i\right)2^{k-2i}}. \tag{11}$$

Applying the multiplication theorem for the gamma function

$$2^{2p}\Gamma(p+\tfrac{1}{2})\Gamma(p+1) = \Gamma(2p+1)\Gamma(1/2),$$

we find

$$E(r^k) = \sum_{i=0}^{[k/2]} \frac{\Gamma(N+k-2i)\Gamma(k+1)\Gamma\frac{(N+2)}{2}\rho^{k-2i}}{\Gamma(N)\Gamma(k-2i+1)\Gamma\left(\frac{N+2}{2}+k-i\right)2^k\Gamma(i+1)}. \tag{12}$$

The above formula may be simplified by considering separately the cases k even and k odd. Setting $2j = k$, (10) becomes

$$E(r^{2j}) = \sum_{i=0}^{j} \frac{\Gamma(N+2j-2i)\Gamma(2j+1)\Gamma\left(\frac{N+2}{2}\right)\rho^{2j-2i}}{\Gamma(N)\Gamma(2j-2i+1)\Gamma\left(\frac{N+2}{2}+2j-i\right)2^{2j}\Gamma(i+1)}. \tag{13}$$

Setting $p = j - i$ and applying the multiplication theorem again, (13) may be written as[2]

$$E(r^{2j}) = \sum_{p=0}^{j} \frac{\Gamma\left(p+\frac{N}{2}\right)\Gamma\left(p+\frac{N+1}{2}\right)\Gamma(j+\frac{1}{2})\Gamma(j+1)\Gamma\frac{(N+2)}{2}\rho^{2p}}{\Gamma\frac{(N)}{2}\Gamma\frac{(N+1)}{2}\Gamma(p+\frac{1}{2})\Gamma(p+1)\Gamma\left(\frac{N+2}{2}+j+p\right)}, \tag{14}$$

[2] For $p = 0$ the expression in the braces $\{\cdots\}$ in (15) is to be taken as 1.

or

$$E(r^{2j}) = \sum_{p=0}^{j} \frac{(2j)!\{N(N+1)(N+2)\cdots(N+2p-1)\}\rho^{2p}}{2^{j-p}(2p)!(j-p)!(N+2)(N+4)\cdots(N+2j+2p)}. \tag{15}$$

The corresponding results for k odd, $k = 2j+1$, are

$$E(r^{2j+1}) = \sum_{p=0}^{j} \frac{\Gamma\left(p+\frac{N+1}{2}\right)\Gamma\left(p+\frac{N+2}{2}\right)\Gamma\left(j+\frac{3}{2}\right)\Gamma(j+1)\left(\frac{N+2}{2}\right)\rho^{2p+1}}{\Gamma\left(\frac{N}{2}\right)\Gamma\left(\frac{N+1}{2}\right)\Gamma(p+1)\Gamma\left(p+\frac{3}{2}\right)\Gamma\left(\frac{N+2}{2}+p+j+1\right)\Gamma(j-p+1)}, \tag{16}$$

$$E(r^{2j+1}) = \sum_{p=0}^{j} \frac{(2j+1)!N(N+1)(N+2)\cdots(N+2p)\rho^{2p+1}}{2^{j-p}(2p+1)!(j-p)!(N+2)(N+4)\cdots(N+2j+2p+2)}. \tag{17}$$

From (15) and (17) we see that

$$\lim_{N\to\infty} E(r^k) = \rho^k, \quad \text{for all } k. \tag{18}$$

3. Specific moments of r. Direct substitution in (15) and (17) yields the following:

$$\begin{aligned}
E(r) &= \frac{N\rho}{N+2} = \mu, \\
E(r^2) &= \frac{1}{N+2} + \frac{N(N+1)\rho^2}{(N+2)(N+4)}, \\
E(r^3) &= \frac{3N\rho}{(N+2)(N+4)} + \frac{N(N+1)(N+2)\rho^3}{(N+2)(N+4)(N+6)}, \\
E(r^4) &= \frac{3}{(N+2)(N+4)} + \frac{6N(N+1)\rho^2}{(N+2)(N+4)(N+6)} \\
&\qquad + \frac{N(N+1)(N+2)(N+3)\rho^4}{(N+2)(N+4)(N+6)(N+8)}.
\end{aligned} \tag{19}$$

The first two moments agree with those obtained by Leipnik, who evaluated them by another method

The central moments of r are

$$\begin{aligned}
E(r-\mu)^2 &= \frac{1}{N+2} - \frac{N(N-2)\rho^2}{(N+2)(N+4)(N+2)} = \sigma^2, \\
E(r-\mu)^3 &= \frac{1}{(N+2)^2}\left(\frac{-6N\rho}{N+4} + \frac{2N(N-2)(3N-2)\rho^3}{(N+2)(N+4)(N+6)}\right) = \mu_3, \\
E(r-\mu)^4 &= \frac{3}{(N+2)(N+4)}\left[1 - \frac{2N(N^2-8N-4)\rho^2}{(N+2)^2(N+6)}\right. \\
&\qquad \left. + \frac{N(N^4-16N^3+40N^2-32N+16)\rho^4}{(N+2)^3(N+6)(N+8)}\right] = \mu_4.
\end{aligned} \tag{20}$$

For large values of N the variance, skewness and kurtosis of r are

$$\sigma^2 = \frac{1-\rho^2}{N+2} + O(N^{-2}),$$
$$\sqrt{\beta_1} = \mu_3/\sigma^3 = o(N^{-1}), \tag{21}$$
$$\beta_2 = \mu_4/\sigma^4 = 3 + o(N^{-1}).$$

These last results are to be expected since it is well known that r has an asymptotic normal distribution.

4. Final remarks. The above results should be adequate, as Leipnik has suggested, for serial correlation problems when $N \geqq 20$. In particular the expressions for the moments of r will be of assistance in evaluating the moments of functions of r; for example, the variance stabilizing transformation $z = \sin^{-1} r$, which will be treated in a future paper.

REFERENCES

[1] R. B. Leipnik, "Distribution of the serial correlation coefficient in a circularly correlated universe," *Ann. Math. Stat.*, Vol. 18 (1947), pp. 80–87.

[2] W. Magnus, and F. Oberhettinger. *Formulas and Theorems for the Special Functions of Mathematical Physics*, Chelsea Publishing Company, New York, 1949.

ON A DECISION PROCEDURE BASED ON THE TUKEY STATISTIC

By K. V. Ramachandran[1] and C. G. Khatri

University of Baroda

1. Summary. In this paper a decision procedure based on the Tukey Studentized range ([5], [6], [8]) has been shown to be an optimum procedure for a particular type of slippage of means of univariate normal populations based on a common but unknown variance. The method given here is similar to that used by Paulson [2] and Truax [7].

2. Introduction. Let $x_{ij}(i = 1, 2, \cdots, k; j = 1, 2, \cdots, n)$ be the elements of k independent samples of size n from normal populations with means μ_i and variance $\sigma^2(i = 1, 2, \cdots, k)$. Let

$$\bar{x}_i = \sum_{j=1}^{n} (x_{ij}/n), \qquad s^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2/k(n-1),$$

$\bar{x}_{\max} = \max(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_k)$ and $\bar{x}_{\min} = \min(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_k)$. Let D_{00} denote the decision that the k means are all equal, and let

$$D_{ij}(i \neq j; i, j = 1, 2, \cdots, k)$$

Received June 11, 1956; revised December 10, 1956.

[1] Present address: Department of Statistics, University of Lucknow, India.

denote the decision that D_{00} is incorrect and $\mu_i = \mu_{\min}$ and $\mu_j = \mu_{\max}$. We will say that the pair (μ_i, μ_j) has slipped by an amount $\Delta(\Delta > 0)$ if $\mu_1 = \mu_2 = \cdots = \mu_{i-1} = \mu_{i+1} = \cdots = \mu_{j-1} = \mu_{j+1} = \cdots = \mu_k = \mu$ (say) and $\mu_i = \mu - \Delta$ and $\mu_j = \mu + \Delta$. The first formulation of the problem is the following: to find a statistical procedure for selecting one of the decisions $(D_{00}, D_{ij})(i \neq j; i, j = 1, 2, \cdots, k)$ which will maximize the probability of making the correct decision when some pair slipped subject to the following restriction (a) when all the means are equal, D_{00} should be selected with probability $1 - \alpha$ (where α is some small positive quantity fixed in advance of the experiment).

Since the class of possible decision procedures seems to be too large to admit an optimum solution we will impose the following restrictions, (b) the decision procedure must be invariant under location and scale transformations of the variates (c) the decision procedure must be symmetric in the sense that the probability of making the correct decision when the pair (μ_i, μ_j) has slipped by an amount Δ must be the same for all $i, j = 1, 2, \cdots, k; i \neq j$. These additional restrictions are rather weak and seem to be reasonable requirements to impose in many practical problems. We will now reformulate the problem as follows: we want a statistical procedure for selecting one of the decisions

$$(D_{00}, D_{ij})(i \neq j; i, j = 1, 2, \cdots, k)$$

which, subject to conditions (a), (b) and (c), will maximize the probability of making the correct decision when one of the pairs has slipped. We shall prove that the optimum solution is the following: if

$$(1) \qquad \bar{x}_i = \bar{x}_{\min}, \quad \bar{x}_j = \bar{x}_{\max}, \quad \text{and} \quad \frac{n(\bar{x}_j - \bar{x}_i)}{[(nk-1)s_0^2]^{1/2}} > q_\alpha,$$

select D_{ij}; if

$$\frac{n(\bar{x}_j - \bar{x}_i)}{[(nk-1)s_0^2]^{1/2}} \leqq q_\alpha,$$

select D_{00}, where q_α is a constant whose value is determined by restriction (a), and $(nk-1)s_0^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2$. This statistic has been suggested, on intuitive grounds, by Tukey [8]. Roy and Bose [6] have shown that the statistic can be derived by the union-intersection principle of test construction. Tables of the distribution of q for different values of α, n, and k are available ([1], [3], [4]).

3. Derivation of the optimum procedure. Since $(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_k, s^2)$ constitute a set of sufficient statistics for the unknown parameters $(\mu_1, \mu_2, \cdots, \mu_k, \sigma^2)$ there is no loss in considering only procedures which depend on this set of statistics. Making use of this in connection with restriction (b) it is easy to see that any allowable decision procedure will depend only on the $k - 1$ statistics $(\bar{x}_i - \bar{x}_k)/s$, $(i = 1, 2, \cdots, k-1)$. Let $w_i = (\bar{x}_i - \bar{x}_k)/s$, $(i = 1, 2, \cdots, k-1)$ and let $\lambda_i = (\mu_i - \mu_k)/\sigma$, $(i = 1, 2, \cdots, k-1)$. The joint distribution

of the set $(w_1, w_2, \cdots, w_{k-1})$ depends only on the parameters $(\lambda_1, \lambda_2, \cdots, \lambda_{k-1})$. Let $\bar{D}_{00}$ be the decision that $\lambda_1 = \lambda_2 = \cdots = \lambda_{k-1} = 0$ and let $\bar{D}_{ij}$ be the decision that $\lambda_1 = \lambda_2 = \cdots = \lambda_{i-1} = \lambda_{i+1} = \cdots = \lambda_{j-1} = \lambda_{j+1} = \cdots = \lambda_{k-1} = 0$, $\lambda_i = -\Delta/\sigma$ and $\lambda_j = \Delta/\sigma$, $(i \neq j;\ i, j = 1, 2, \cdots, k-1)$, while $\bar{D}_{ik}$ denotes the decision that $\lambda_1 = \lambda_2 = \cdots = \lambda_{i-1} = \lambda_{i+1} = \cdots = \lambda_{k-1} = -\Delta/\sigma$, $\lambda_i = -2\Delta/\sigma$, $(i = 1, 2, \cdots, k-1)$ and $\bar{D}_{ki}$ denotes the decision that $\lambda_1 = \lambda_2 = \cdots = \lambda_{i-1} = \lambda_{i+1} = \cdots = \lambda_{k-1} = \Delta/\sigma$, $\lambda_i = 2\Delta/\sigma$ $(i = 1, 2, \cdots, k-1)$. Since any allowable decision procedure for selecting one of the set (D_{00}, D_{ij}) $(i \neq j;\ i, j = 1, 2, \cdots, k)$ must be a function only of $(w_1, w_2, \cdots, w_{k-1})$, it can be transformed into a procedure for selecting one of the set $(\bar{D}_{00}, \bar{D}_{ij})$, $(i \neq j,\ i, j = 1, 2, \cdots, k)$ by making D_{ij} correspond to $\bar{D}_{ij}$, $(i, j = 0, 1, 2, \cdots, k;\ i \neq j$ if $i, j > 0)$, i.e. whenever the original decision procedure selects D_{ij}, the transformed decision procedure is to select $\bar{D}_{ij}$. Because of restriction (a), the probability that any transformed allowable decision procedure will select D_{00} when $\lambda_1 = \lambda_2 = \cdots = \lambda_{k-1} = 0$ will be equal to $1 - \alpha$, in addition the probability that any allowable decision procedure will select D_{ij} when the pair (μ_i, μ_j) has slipped is equal to the probability that the transformed procedure select $\bar{D}_{ij}$ when $\bar{D}_{ij}$ is the correct decision, and this last probability must be the same for each (i, j) because of restriction (c).

The proof that (1) is the optimum solution consists mainly in showing that for any Δ and σ there exist a set of nonzero a priori probabilities g_{00}, g_{ij}, $(i \neq j;\ i, j = 1, 2, \cdots, k)$ which are functions of Δ and σ so that when (1) is transformed in the manner indicated above into a decision procedure for selecting one of $(\bar{D}_{00}, \bar{D}_{ij})$, $(i \neq j;\ i, j = 1, 2, \cdots, k)$, it will maximize the probability of making the correct decision among the set $(\bar{D}_{00}, \bar{D}_{ij})$, $(i \neq j;\ i, j = 1, 2, \cdots, k)$ when g_{ij} is the probability that $\bar{D}_{ij}$ is the correct decision $(i, j = 0, 1, 2, \cdots, k, i \neq j$ if $i, j > 0)$.

Assuming this has been demonstrated, it follows easily that (1) must be the optimum solution. For suppose there existed an allowable decision procedure D^*, which for some Δ and σ had a greater probability than (1) of making the correct decision when some pair had slipped. Then D^*, which must only be a function of $(w_1, w_2, \cdots, w_{k-1})$ when transformed in the indicated manner into a decision procedure for selecting one of $(\bar{D}_{00}, \bar{D}_{ij})$, $(i \neq j, i, j = 1, 2, \cdots, k)$ will have greater probability than (1) of making the correct decision among $(\bar{D}_{00}, \bar{D}_{ij})$, $(i \neq j;\ i, j = 1, 2, \cdots, k)$ with respect to any set of nonzero a priori probabilities, which would be a contradiction.

To show that the required a priori distribution exists, first let $u_i = (\bar{x}_i - \bar{x}_k)/\sigma$, $(i = 1, 2, \cdots, k-1)$ so that $w_i = u_i\sigma/s$.

The joint probability density function $f(w_1, w_2, \cdots, w_{k-1})$ of $w_1, w_2, \cdots, w_{k-1}$ is easily found to be given by

$$f(w_1, w_2, \cdots, w_{k-1}) = C \int_0^\infty t^{n'+k-2} \exp \tag{2}$$
$$- \frac{1}{2}\left[n't^2 + A \sum_{i=1}^{k-1} (w_i t - \lambda_i)^2 + B \sum_{i \neq j} (w_i t - \lambda_i)(w_j t - \lambda_j) \right] dt,$$

where $n' = k(n-1)$, $A = n(k-1/k)$, $B = -n/k$, and C is a constant whose precise value is not needed.

Let $f_{ij} = f(w_1, w_2, \cdots, w_{k-1} | \bar{D}_{ij})$, $(i, j = 0, 1, 2, \cdots, k; i \neq j$ if $i, j > 0)$ be the joint probability density function of $w_1, w_2, \cdots, w_{k-1}$ when $\bar{D}_{ij}$ is the correct decision. The decision procedure which will maximize the probability of making the correct decision among the set $(\bar{D}_{00}, \bar{D}_{ij})$, $(i \neq j; i, j = 1, 2, \cdots, k)$ when the a priori probability distribution is $(p_{00}, p_{12}, \cdots, p_{k-1k})$, i.e. the Bayes solution with respect to $(p_{00}, p_{12}, \cdots, p_{k-1k})$, is known [9] to be given by the rule: for each i, j $(i, j = 0, 1, 2, \cdots, k; i \neq j$ if $i, j > 0)$ select $\bar{D}_{ij}$ for all points in the w space where $p_{ij}f_{ij} = \max (p_{00}f_{00}, p_{12}f_{12}, \cdots, p_{k-1k}f_{k-1k})$. For the problem at hand, this is the unique Bayes solution except possibly for a set of measure zero according to all f_{ij}. Using (2) it is easy to calculate for each i, j the region where $\bar{D}_{ij}$ is selected for the special a priori distribution $p_{00} = 1 - k(k-1)p$, $p_{12} = \cdots = p_{k-1k} = p$.

It can be easily checked that the Bayes solution is the following procedure: Select $\bar{D}_{ij} (i, j = 1, 2, \cdots, k-1, i \neq j)$ if

$$(w_j - w_i) > (w_{j'} - w_{i'}) \cdot (i', j' = 1, 2, \cdots, k-1, i' \neq j' \quad \text{and} \quad i \neq i', j \neq j' \quad \text{simultaneously})$$

$$(w_j - w_i) > |w_{i'}| \quad \text{and} \quad \frac{(w_j - w_i)(A - B)}{\sqrt{n' + A\sum_{r=1}^{k-1} w_r^2 + B \sum_{r \neq s=1}^{k-1} w_r w_s}} > q; \tag{3}$$

Select $\bar{D}_{ik} (i = 1, 2, \cdots, k-1)$ if

$$-w_i > (w_{j'} - w_{i'})(i', j' = 1, 2, \cdots k-1, i' \neq j')$$

$$\text{and} \quad \frac{-w_i(A - B)}{\sqrt{n' + A\sum_{r=1}^{k-1} w_r^2 + B \sum_{r \neq s=1}^{k-1} w_r w_s}} > q; \tag{4}$$

Select $\bar{D}_{ki} (i = 1, 2, \cdots, k-1)$ if

$$w_i > (w_{j'} - w_{i'})(i', j' = 1, 2, \cdots, k-1, i' \neq j')$$

$$\text{and} \quad \frac{w_i(A - B)}{\sqrt{n' + A\sum_{r=1}^{k-1} w_r^2 + B \sum_{r \neq s=1}^{k-1} w_r w_s}} > q; \tag{5}$$

and select $\bar{D}_{00}$ otherwise.

Define the function $F(p)$ by the equation

$$F(p) = \int_0^\infty y^{n'+k-2} \exp\left(\frac{-y^2}{2}\right) \left\{p \exp\left(\frac{-n\Delta^2}{2}\right) \exp\left(\frac{\Delta y q_\alpha}{\sigma}\right) - [1 - k(k-1)p]\right\} dy, \tag{6}$$

where q_α is the constant used in (1).

It is obvious that $F(p)$ is a continuous function of p with $F(0) < 0$ and $F(1/k(k-1)) > 0$. Hence there exists a p^* with $0 < p^* < 1/k(k-1)$ which is a function of Δ/σ so that $F(p^*) = 0$. Once the Bayes solution relative to $[1 - k(k-1)p, p, \cdots, p]$ has been worked out, it is obvious that to get the Bayes solution relative to $[1 - k(k-1)p^*, p^*, \cdots, p^*]$ it is only necessary to replace q by q_a. If we now substitute $w_i = (\bar{x}_i - \bar{x}_k)/s$ and replace A and B by their values, we find after some simplifications that the Baye's solution relative to $[1 - k(k-1)p^*, p^*, \cdots, p^*]$ reduces to (1) when $\bar{D}_{ij}$ is made to correspond to D_{ij}, $(i, j = 0, 1, 2, \cdots, k; i \neq j$ if $i, j > 0)$. Since (1) is an allowable procedure, this proves that it is an optimum one.

REFERENCES

[1] Joyce M. May, "Extended and corrected tables of the upper percentage points of the Studentized range," *Biometrika*, Vol. 39 (1952), pp. 192–193.

[2] Edward Paulson, "An optimum solution to the k sample slippage problem for the normal distribution," *Annals of Math. Stat.*, Vol. 23 (1952), pp. 610–616.

[3] E. S. Pearson and H. O. Hartley, "Biometrika Tables for Statisticians," (Cambridge University Press), Vol. 1 (1954).

[4] K. C. S. Pillai, "On the distribution of Studentized range," *Biometrika*, Vol. 39 (1952), pp. 194–195.

[5] K. V. Ramachandran, "On the Tukey test for the equality of means and the Hartley test for the equality of variances," *Annals of Math. Stat.*, Vol. 27 (1956), pp. 825–831.

[6] S. N. Roy and R. C. Bose, "Simultaneous confidence internal estimation," *Annals of Math. Stat.*, Vol. 24 (1953), pp. 513–536.

[7] Donald R. Truax, "An optimum slippage test for the variances of k normal distributions," *Annals of Math. Stat.*, Vol. 24 (1953), pp. 669–674.

[8] J. W. Tukey, "Allowances for various types of error rates," (unpublished invited address, Blacksburg meeting of the institute of mathematical statistics, March, 1952).

[9] A. Wald, "Statistical decision functions," John Wiley and Sons, 1950.

ESTIMATES OF THE MEAN AND STANDARD DEVIATION OF A NORMAL POPULATION[1]

By W. J. Dixon

University of California, Los Angeles

0. Summary. Several simple estimates of the mean and standard deviation of a normal population are discussed. The efficiencies of these estimates are compared to the sample mean and sample standard deviation and to the best linear unbiased estimates. Little efficiency is lost when simple rather than optimum weights are used.

Received October 29, 1956; revised February 5, 1957.

[1] This research sponsored in part by the Office of Naval Research. It may be reproduced in whole or in part for any purpose of the United States Government.

TABLE I

Several Estimates of Mean of Normal Population with Efficiencies. (Variances to be multiplied by σ^2.)

N	Median		Midrange		$(X_i + X_j)/2$			$\bar{X}_{)1,N(}$		
	Var.	Eff.	Var.	Eff.	i, j	Var.	Eff.	Var.	Eff.	A^*
2	0.500	1.00	0.500	1.00	1, 2	0.500	1.00			
3	0.449	0.743	0.362	0.920	1, 3	0.362	0.920	0.449	0.743	1.000
4	0.298	0.838	0.298	0.838	2, 3	0.298	0.838	0.298	0.838	1.000
5	0.287	0.697	0.261	0.767	2, 4	0.231	0.867	0.227	0.881	0.994
6	0.215	0.776	0.236	0.706	2, 5	0.193	0.865	0.184	0.906	0.992
7	0.210	0.679	0.218	0.654	2, 6	0.168	0.849	0.155	0.922	0.990
8	0.168	0.743	0.205	0.610	3, 6	0.149	0.837	0.134	0.934	0.990
9	0.166	0.669	0.194	0.572	3, 7	0.132	0.843	0.118	0.942	0.990
10	0.138	0.723	0.186	0.539	3, 8	0.119	0.840	0.105	0.949	0.990
11	0.137	0.663	0.178	0.510	3, 9	0.109	0.832	0.0952	0.955	0.991
12	0.118	0.709	0.172	0.484	4, 9	0.100	0.831	0.0869	0.959	0.991
13	0.117	0.659	0.167	0.461	4, 10	0.0924	0.833	0.0799	0.963	0.991
14	0.102	0.699	0.162	0.440	4, 11	0.0860	0.830	0.0739	0.966	0.992
15	0.102	0.656	0.158	0.422	4, 12	0.0808	0.825	0.0688	0.969	0.992
16	0.0904	0.692	0.154	0.392	5, 12	0.0756	0.827	0.0644	0.971	0.993
17	0.0901	0.653	0.151	0.389	5, 13	0.0711	0.827	0.0605	0.973	0.993
18	0.0810	0.686	0.148	0.375	5, 14	0.0673	0.825	0.0570	0.975	0.993
19	0.0808	0.651	0.145	0.363	6, 14	0.0640	0.823	0.0539	0.976	0.993
20	0.0734	0.681	0.143	0.350	6, 15	0.0607	0.824	0.0511	0.978	0.994
∞		0.637		0.000	0.27, 0.73		0.810		1.000	1.000

* A = Var (BLSS)/Var$(\bar{X}_{)1,N(})$.

Since moments of the order statistics are now available for samples of sizes $N \leqq 20$ from normal populations [3] it is a simple matter to find the variances of linear combinations of order statistics. The sample values are denoted $X_1 \leqq X_2 \leqq X_3 \leqq \cdots \leqq X_N$.

1. Estimates of the mean. Table I gives the variance and efficiency of the following estimates of the population mean: (a) median, (b) midrange, (c) mean of best two, and (d) $\bar{X}_{)1,N(} = \sum_{i=2}^{N-1} X_i / (N-2)$. The median and midrange are given primarily for comparison purposes, since results are well known. The median is defined as $X_{(N+1)/2}$ for N odd and as $\frac{1}{2}(X_{N/2} + X_{(N+1)/2})$ for N even. The mean of the best two (here "best" is used in the sense of minimum variance) is reported as the small sample equivalent of the estimate commonly used in large samples, the mean of the 27th and 73rd percentiles. It can be seen from Table I that for sample sizes larger than five, the particular ordered observations indicated are not far from the 27th and 73rd percentiles, and efficiencies are close to the asymptotic efficiency (0.810). The efficiency reported for the mean, median, and the mean of best two is the ratio of the variance of the statistic to the variance of the arithmetic mean.

Estimate (d) above is the mean of all observations except the largest and

TABLE II

A linear estimate of the standard deviation. (Variances to be multiplied by σ^2.)

Sample Size	Range			s'			
	k	Var.	Eff.	Estimate	Var.	Eff.	B^*
2	0.886	0.571	1.000	$0.8862w$	0.571	1.000	1.000
3	0.591	0.275	0.992	$0.5908w$	0.275	0.992	1.000
4	0.486	0.183	0.975	$0.4857w$	0.183	0.975	0.986
5	0.430	0.138	0.955	$0.4299w$	0.138	0.955	0.966
6	0.395	0.112	0.933	$0.2619(w + w_{(2)})$	0.109	0.957	0.968
7	0.370	0.0949	0.911	$0.2370(w + w_{(2)})$	0.0895	0.967	0.978
8	0.351	0.0829	0.890	$0.2197(w + w_{(2)})$	0.0761	0.970	0.980
9	0.337	0.0740	0.869	$0.2068(w + w_{(2)})$	0.0664	0.968	0.979
10	0.325	0.0671	0.850	$0.1968(w + w_{(2)})$	0.0591	0.964	0.974
11	0.315	0.0616	0.831	$0.1608(w + w_{(2)} + w_{(4)})$	0.0529	0.967	0.977
12	0.307	0.0571	0.814	$0.1524(w + w_{(2)} + w_{(4)})$	0.0478	0.972	0.981
13	0.300	0.0533	0.797	$0.1456(w + w_{(2)} + w_{(4)})$	0.0436	0.975	0.984
14	0.294	0.0502	0.781	$0.1399(w + w_{(2)} + w_{(4)})$	0.0401	0.977	0.985
15	0.288	0.0474	0.766	$0.1352(w + w_{(2)} + w_{(4)})$	0.0372	0.977	0.985
16	0.283	0.0451	0.751	$0.1311(w + w_{(2)} + w_{(4)})$	0.0347	0.975	0.983
17	0.279	0.0430	0.738	$0.1050(w + w_{(2)} + w_{(3)} + w_{(5)})$	0.0325	0.978	0.985
18	0.275	0.0412	0.725	$0.1020(w + w_{(2)} + w_{(3)} + w_{(5)})$	0.0305	0.978	0.986
19	0.271	0.0395	0.712	$0.09939(w + w_{(2)} + w_{(3)} + w_{(5)})$	0.0288	0.979	0.986
20	0.268	0.0381	0.700	$0.10446(w + w_{(3)} + w_{(4)} + w_{(5)})$	0.0272	0.980	0.987

* B = Var (BLSS)/Var(s').

smallest. Interest in this statistic arises when the extreme observations are poorly determined or not available. References [1] and [2] refer to this condition as doubly censored and develop best linear systematic statistics (BLSS) for various amounts of single and double censoring of the sample. The decrease in efficiency of the simple unweighted mean $\bar{X}_{)1,N(}$ compared to the BLSS based on the same observations is not great. In no case is the loss in efficiency more than 1 per cent. This can be noted from the ratio Var(BLSS) / Var($\bar{X}_{)1,N(}$) given in Table I, since this ratio is never less than 0.990. It seems likely that for many applications, one could dispense with the use of unequal weights for the systematic statistics in this case. It can be seen that the efficiency is not greatly affected by the use of coefficients differing greatly from the optimum. The column head "Eff." is efficiency of $\bar{X}_{)1,N(}$ compared with the mean of all observations and is approximately the same for the BLSS.

2. Estimates of standard deviation. The efficiency of the range, w, as an estimate of the standard deviation in small samples, is well known. Similar estimates using additional observations will also give high efficiencies for larger sample sizes. Table II contains the efficiency of the range estimate compared to the unbiased estimate based on the sample standard deviation. The quantity k which satisfies $E(kw) = \sigma$ is tabled for reference. Let us denote the subranges

$X_{N-i+1} - X_i$ by $w_{(i)}$ and $w_{(1)} = w$. The unbiased estimate of the type $s' = k'(\sum w_{(i)})$, where the summation is over the subset of all $w_{(i)}$ which gives minimum variance, is indicated in Table II. The column headed "Eff." refers to the comparison with the unbiased sample standard deviation. The final column gives the ratio of the variance of the best linear systematic statistic as given in [2] to the variance of s'. By examining this ratio we can see that the loss in efficiency due to the use of "zero or one" weights for each range rather than the optimum weights given in [2], is not great.

REFERENCES

[1] A. K. Gupta, "Estimation of the mean and standard deviation of a normal population from a censored sample," *Biometrika*, Vol. 39 (1952), pp. 260–273.

[2] A. E. Sarhan and B. G. Greenberg, "Estimation of location and scale parameters by order statistics from singly and doubly censored samples," Part 1, *Ann. Math. Stat.*, Vol. 27 (1956), pp. 427–451.

[3] Dan Teichroew, "Tables of expected values of order statistics and products of order statistics for samples of size twenty and less from the normal distribution," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 410–426.

THE INDIVIDUAL ERGODIC THEOREM OF INFORMATION THEORY[1]

By Leo Breiman

University of California, Berkeley

1. Introduction. Information theory is largely concerned with stationary stochastic processes $\cdots x_{-1}, x_0, x_1, \cdots$ taking values in a finite "alphabet," $a_1, \cdots, a_s$. In addition, it is usually assumed that the processes are ergodic, that is to say, the shift operator T, defined on the sequence space Ω of the process by shifting each coordinate of a sequence once to the right, is metrically transitive with respect to the probability measure p on Ω.

A question of importance in information theory regarding these processes is the nature and existence, in some sense, of the expression

$$\text{(a)} \qquad \lim_n \left(-\frac{1}{n} \log_2 p(x_0, \cdots, x_{n-1}) \right).$$

In 1948 Shannon [1] showed that for stationary, ergodic Markov chains (a) exists as a limit in probability and is equal to a constant. This limiting constant was termed by Shannon the "entropy" of the process. In 1953 McMillan [2] lifted the restriction to Markov chains and proved that if the process is merely stationary and ergodic, then (a) exists as a limit in L_1 mean and is constant. The purpose of this note is to prove that under the same conditions the limit (a) exists almost surely (a.s.).

Received October 15, 1956.

[1] This paper was prepared with the support of the Office of Ordnance Research, U. S. Army under Contract DA-04-200-ORD-171.

2. The modified Birkhoff theorem. The heart of the matter is the following modification of the individual ergodic theorem.

THEOREM 1. *Let T be a metrically transitive $1-1$ measure preserving transformation of the probability space $(\Omega, \mathcal{B}, p)$ onto itself. Let $g_0(\omega), g_1(\omega), \cdots$ be a sequence of measurable functions on Ω converging* a.s. *to the function $g(\omega)$ such that $E(\sup_k |g_k|) < \infty$. Then*

$$\lim_n \frac{1}{n}\sum_{k=0}^{n-1} g_k(T^k\omega) = Eg \text{ a.s.}$$

Proof. We write

$$\frac{1}{n}\sum_{k=0}^{n-1} g_k(T^k\omega) = \frac{1}{n}\sum_{k=0}^{n-1} g(T^k\omega) + \frac{1}{n}\sum_{k=0}^{n-1}[g_k(T^k\omega) - g(T^k\omega)].$$

The conditions of the theorem imply that $E|g| < \infty$ and by Birkhoff's ergodic theorem (see, for example, [3], pp. 464–469), the first term on the right above converges a.s. to Eg. It remains to show that the second term converges a.s. to zero. Let $G_N(\omega) = \sup_{k \geq N} |g_k(\omega) - g(\omega)|$, then for every fixed N

$$\overline{\lim}\left|\frac{1}{n}\sum_{k=0}^{n-1}[g_k(T^k\omega) - g(T^k\omega)]\right| \leqq \overline{\lim}\frac{1}{n}\sum_{k=0}^{n-1}|g_k(T^k\omega) - g(T^k\omega)|$$

$$\leqq \overline{\lim}\frac{1}{n}\sum_{k=0}^{n-1} G_N(T^k\omega) = EG_N \quad \text{a.s.}$$

The sequence $\{G_N\}$ converges monotonically to zero and

$$EG_0 \leqq E(\sup_k |g_k| + |g|) < \infty,$$

so by the monotone convergence theorem $EG_N \to 0$, which proves the theorem.

THEOREM 2. *Let $\cdots, x_{-1}, x_0, x_1, \cdots$ be a stationary ergodic process ranging over a finite number of values $a_1, \cdots, a_s$. Then there is a constant H such that*

$$\lim_n\left(-\frac{1}{n}\log_2 p(x_0, \cdots, x_{n-1})\right) = H \quad \text{a.s.}$$

Proof. Let

$$g_0(\omega) = -\log_2 p(x_0),$$

$$g_k(\omega) = -\log_2 \frac{p(x_{-k}, x_{-k+1}, \cdots, x_0)}{p(x_{-k}, x_{-k+1}, \cdots, x_1)}, \qquad k \geqq 1.$$

Then, letting T be the shift operator,

$$-\frac{1}{n}\log_2 p(x_0, \cdots, x_{n-1}) = \frac{1}{n}\sum_{k=0}^{n-1} g_k(T^k\omega).$$

Since T is $1-1$, measure preserving and metrically transitive, we apply Theorem 1 and our work will be done as soon as we show that the sequence $\{g_k\}$ converges a.s. and that $E(\sup_k g_k) < \infty$.

To do this we use the inequality established by McMillan [2],

$$\text{(i)} \qquad \int_{[m \leq g_k < m+1]} g_k \leqq s(m+1)2^{-m}.$$

We confine our attention to the cylinder set $Z_i \subset \Omega$, $Z_i = \{\omega; x_0 = a_i\}$. On Z_i we have

$$g_k(\omega) = -\log_2 p(x_0 = a_i \mid x_{-1}, \cdots, x_{-k}).$$

Since $p(x_0 = a_i \mid x_{-1}, \cdots, x_{-k})$ is a martingale, it follows from the convexity of $-\log$ and inequality (i) that the sequence $\{g_k\}$ is a semi-martingale (see [3], p. 295). Therefore, g_k converges a.s. on Z_i and hence on Ω.

Furthermore, by a semi-martingale inequality, [3] p. 317, we have, on Z_i,

$$\int_{Z_i} \left(\sup_{0 \leq k \leq n} g_k \right) \leqq \frac{e}{e-1} + \frac{e}{e-1} \int_{Z_i} (g_n \log^+ g_n).$$

By using inequality (i) again, we bound the last term on the above right;

$$\int_{Z_i} (g_n \log^+ g_n) = \sum_{m=0}^{\infty} \int_{Z_i [m \leq g_n < m+1]} (g_n \log^+ g_n)$$

$$\leqq \sum_{m=0}^{\infty} s(m+1) \log (m+1) 2^{-m}.$$

Therefore $\int_{Z_i} (\sup_k g_k) < \infty$, by addition $E(\sup_k g_k) < \infty$, and the theorem is proved.

It is a pleasure to acknowledge our debt to Professor David Blackwell who suggested to us the problem treated herein.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, Vol. 27 (1948), pp. 379–423, pp. 623–656.

[2] B. McMillan, "The basic theorems of information theory," *Ann. Math. Stat.*, Vol. 24 (1953), pp. 196–219.

3] J. L. Doob, *Stochastic Processes*, John Wiley & Sons, Inc., New York, 1953.

A COUNTEREXAMPLE TO A THEOREM OF KOLMOGOROV[1,2]

By Leo Breiman

University of California, Berkeley

1. Introduction. In 1928 Kolmogorov [1] presented the now well-known degenerate convergence theorem (weak law of large numbers) as follows (see, for

Received August 10, 1956; revised January 18, 1957.

[1] This paper was prepared with the support of the Office of Ordnance Research, U. S. Army under Contract DA-04-200-ORD-171.

[2] After this note was submitted the author was informed that C. Derman had constructed a similar counterexample. While the note was in proof, a similar counterexample appeared in a paper by Hartley Rogers, Jr., *Proc. Am. Math. Soc.*, Vol. 8 (1957), pp. 518–520.

example, Loève [2]): let $X_1, X_2, \cdots$ be independent random variables such that $EX_k = 0$, $k = 1, 2, \cdots$, and let

$$X_{nk} = \begin{cases} X_k & \text{if } |X_k| < n, \\ 0 & \text{if } |X_k| \geqq n. \end{cases}$$

Then

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{P} 0$$

if and only if

(i) $\sum_{k=1}^{n} P(|X_k| \geqq n) \to 0$,

(ii) $\frac{1}{n}\sum_{k=1}^{n} EX_{nk} \to 0$,

(iii) $\frac{1}{n^2}\sum_{k=1}^{n} \sigma^2 X_{nk} \to 0$.

Presented without proof in the same paper was a sharpened version of the above theorem with condition (iii) replaced by

(iii′) $\frac{1}{n^2}\sum_{k=1}^{n} EX_{nk}^2 \to 0$.

A proof of this last theorem was given in Gnedenko and Kolmogorov's 1949 book and carried over into the English edition ([3], pp. 135–137). Unfortunately, the proof contains a slight gap and the sharpened theorem is not correct. Since it appears in several places in the literature, for example, in Loève ([2], p. 278), and follows from Theorems 3.2 (p. 124) and 3.3 (p. 125) of Doob's book [4] the following simple counterexample may be of interest to the reader:

We will show that conditions (i), (ii), (iii′) are not necessary by proceeding as follows: define the independent random variables $X_1, X_2, \cdots$ by

$$\left.\begin{aligned} P(X_1 = 0) &= 1, \\ P(X_k = (-1)^k k^{5/2}) &= k^{-2}, \\ P(X_k = (-1)^{k+1} k^{1/2}(1 - k^{-2})^{-1}) &= 1 - k^{-2}, \end{aligned}\right\} \qquad k \geqq 2.$$

We verify immediately that $EX_k = 0$, $k = 1, 2, \cdots$. Then we demonstrate that conditions (i), (ii), (iii) above are satisfied. Finally we show that, contrary to the theorem,

$$\frac{1}{n^2}\sum_{k=1}^{n} EX_{nk}^2 \nrightarrow 0.$$

In the following proofs we take $n \geqq 4$.

Proof of (i). If $k^{5/2} < n$, then $X_{nk} = X_k$, and if $k \leqq n$, then $k^{1/2}(1 - k^{-2})^{-1} < n$. Hence

$$P(|X_k| \geqq n) = \begin{cases} 0 & \text{if } 1 \leqq k^{5/2} < n, \\ k^{-2} & \text{if } n \leqq k^{5/2} \text{ and } k \leqq n, \end{cases}$$

and

$$\sum_{k=1}^{n} P(|X_k| \geqq n) = \sum_{k=[n^{3/5}]}^{n} \frac{1}{k^2} \to 0,$$

where $[\cdot]$ denotes next higher integer.

PROOF OF (ii).

$$EX_{nk} = \begin{cases} 0 & \text{if } 1 \leqq k^{5/2} < n \\ (-1)^{k+1}k^{1/2} & \text{if } n \leqq k^{5/2} \text{ and } k \leqq n. \end{cases}$$

Hence

$$\frac{1}{n}\sum_{k=1}^{n} EX_{nk} = \frac{1}{n}\sum_{k=[n^{3/5}]}^{n} (-1)^{k+1}k^{1/2}.$$

We use the inequality, valid for $s > 1$,

$$\sqrt{s} - \sqrt{s-1} < \frac{1}{2\sqrt{s-1}}$$

to get

$$\left|\frac{1}{n}\sum_{k=1}^{n} EX_{nk}\right| \leqq \left(\frac{1}{2n}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k}} + \frac{1}{n}\sqrt{[n^{2/5}]}\right) \to 0.$$

PROOF OF (iii).

$$\sigma^2 X_{nk} = \begin{cases} k^3 + k(1-k^{-2})^{-1} & \text{if } 2 \leqq k^{5/2} < n, \\ k(1-k^{-2})^{-1} - k & \text{if } n \leqq k^{5/2} \text{ and } k \leqq n. \end{cases}$$

For $k \geqq 2$,

$$k^3 + k(1-k^{-2})^{-1} \leqq k^3 + \tfrac{4}{3}k \leqq 2k^3,$$

$$k(1-k^{-2})^{-1} - k = k^{-1}(1-k^{-2})^{-1} \leqq \tfrac{4}{3}k^{-1}.$$

Hence

$$\frac{1}{n^2}\sum_{k=2}^{n}\sigma^2 X_{nk} \leqq \frac{1}{n^2}\sum_{k=2}^{[n^{3/5}]} 2k^3 + \frac{1}{n^2}\sum_{k=[n^{3/5}]}^{n}\frac{4}{3k} \leqq \frac{1}{2n^2}([n^{2/5}])^4 + \frac{4}{3n^2}\log(n) \to 0.$$

Finally, we show that $1/n^2\left(\sum_{k=1}^{n} EX_{nk}^2\right) \nrightarrow 0$. We have

$$\frac{1}{n^2}\sum_{k=1}^{n} EX_{nk}^2 \geqq \frac{1}{n^2}\sum_{k=[n^{3/5}]}^{n} EX_{nk}^2 = \frac{1}{n^2}\sum_{k=[n^{3/5}]}^{n} k(1-k^{-2})^{-1} \geqq \frac{1}{n^2}\sum_{k=[n^{3/5}]}^{n} k \to \tfrac{1}{2},$$

which completes the counterexample.

It is a pleasure to be able to acknowledge our debt to M. Loève, who brought the question to our attention and suggested further inquiry. We are also indebted to R. K. N. Patell whose letter to M. Loève was the cause of the reexamination of this theorem.

REFERENCES

[1] A. KOLMOGOROV, "Ueber die Summen durch den Zufall bestimmter unabhangigen Grossen," *Math. Ann.*, Vol. 99 (1928), pp. 309-319.
[2] M. LOÈVE, *Probability Theory*, Van Nostrand, New York, 1955.
[3] B. GNEDENKO AND A. KOLMOGOROV, *Limit Distributions for Sums of Independent Random Variables*, Addison-Wesley, Cambridge, 1954.
[4] J. L. DOOB, *Stochastic Processes*, John Wiley & Sons, New York, 1953.

ON THE COMBINING OF INTERBLOCK AND INTRABLOCK ESTIMATES

BY D. A. S. FRASER

University of Toronto

In a recent paper Sprott [1] has considered methods for combining interblock and intrablock estimates of variety contrasts for incomplete block designs. The intrablock estimates are derived from treatment contrasts obtained within blocks. The interblock estimates presuppose that the block effects are random, independent, and identically distributed, and they are derived from contrasts among the block averages. Under normality the intrablock estimates are independent of the interblock estimates.

Sprott compares two methods for producing combined estimates. The first method, introduced by Yates [2], is the familiar method of combining by weighting with the reciprocal of the variances, and is known to produce minimum variance when two real estimates of the same quantity are combined linearly. The second method, discussed by Rao [3] and Cochran and Cox [4], is to apply the method of maximum likelihood to the joint density function, and the resulting estimate is linear in terms of the interblock and intrablock estimates. Sprott shows that, in general, the two methods are not equivalent. The second method is direct and has considerable theoretical weight behind its use. We are left then with the implication that one of the methods is incorrect for obtaining good estimates. In a sense this is not the case. Rather, one of the methods may be *inappropriately* applied. Weighting with reciprocal variances is *appropriate* to combining real estimates but if applied to vector estimates it ignores any covariances and may not be optimum.

Suppose $x = (x_1, \cdots, x_r)$ and $y = (y_1, \cdots, y_r)$ are independent estimates of the parameter $\eta = (\eta_1, \cdots, \eta_r)$ and have nonsingular covariance matrices V and W respectively. Also suppose, for the moment, that x and y are normal. Then, the joint density function is a constant times

$$\exp\left[-\tfrac{1}{2}(x-\eta)V^{-1}(x-\eta)' - \tfrac{1}{2}(y-\eta)W^{-1}(y-\eta)'\right],$$

Received October 29, 1956; revised January 10, 1957.

which can be factored as

$$\exp\{[(xV^{-1}+yW^{-1})(V^{-1}+W^{-1})^{-1}-\eta](V^{-1}+W^{-1})[(xV^{-1}+yW^{-1})(V^{-1}+W^{-1})^{-1}-\eta]'\} \times \exp\{(x-y)(V+W)^{-1}(x-y)'\}. \tag{1}$$

Then, assuming that the covariances are known, we see that

$$(xV^{-1}+yW^{-1})(V^{-1}+W^{-1})^{-1} \tag{2}$$

is a sufficient statistic for η. Also, from its non-singular distribution, it follows that (2) is a complete statistic. Hence, by the theorems of Lehmann and Scheffé which may be found on pp. 61–64 in [5], an unbiased estimate of η based on (2) will be unique, and it will have minimum concentration ellipsoid among *all* unbiased estimates, linear or not. Each coordinate will also have minimum variance. From (1), we see immediately that (2), as it stands, is an unbiased estimate of η, and hence has the properties above. If the assumption of normality is removed, (2) remains unbiased and has minimum variances among unbiased linear estimates.

The question arises as to when the two methods are equivalent and the more-direct first method usable in place of the second. A combined estimate based on the first method would have the form

$$xD + y(I - D), \tag{3}$$

where D is a diagonal matrix and I is the identity matrix ($r \times r$). If (2) reduces to this form (3), then

$$V^{-1}(V^{-1}+W^{-1})^{-1} = (I + W^{-1}V)^{-1} \tag{4}$$

is diagonal and hence VW^{-1} is diagonal: $VW^{-1} = D^*$. We have

$$V = D^*W = WD^*, \tag{5}$$

where the second expression follows from the symmetry of covariance matrices. (5) implies that for any non-zero off-diagonal element in W the corresponding two coordinate indices have equal diagonal elements in D^*. Therefore, by permuting coordinates it follows that W can be made diagonal in blocks and that V is obtained by multiplying each block by a positive constant. Thus for the vectors x, y the rearranged coordinates fall into independent groups. A group for x and a group for y have the same covariance matrix except for a *single* scale factor. There are two extremes for this: first, the x and y have the same covariance matrix except for a single factor; second, the x and y both have independent coordinates.

Thus, in general, it is not enough to combine estimates, coordinate by coordinate. An estimate with optimum properties is obtained by weighting the *vectors* by the inverse of their covariance matrices. We can easily see that doing this for the incomplete block problem considered by Sprott will produce the estimates as obtained by the second method. For, the second method uses maxi-

mum likelihood on the joint density, and from (1) we see that this obviously produces (2).

REFERENCES

[1] D. A. SPROTT, "A Note on combined interblock and intrablock estimation in incomplete block designs," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 633–641.

[2] F. YATES, "The recovery of interblock information in balanced incomplete block designs," *Ann. Eugenics*, Vol. 10 (1940), pp. 317–325.

[3] C. R. RAO, "General methods of analysis for incomplete block designs," *J. Amer. Stat. Assn.*, Vol. 42 (1947), pp. 541–561.

[4] W. G. COCHRAN AND G. M. COX, *Experimental Designs*, John Wiley and Sons, New York, 1950.

[5] D. A. S. FRASER, *Nonparametric Methods in Statistics*, John Wiley and Sons, New York, 1957.

ABSTRACTS OF PAPERS

(Additional abstracts of papers presented at the Washington meeting of the Institute, March 7–9, 1957)

1. The K-Visit Method of Consumer Testing, GEORGE E. FERRIS, General Foods Corporation, (By Title).

When testing a pair of products for consumer preference the problem of how to treat or interpret no preference votes arises. A method is described of collecting data from a given number of consumers by repeated visits to them, or by obtaining repeated judgments from them in stores, which enables the estimation of the true proportion of those consumers in the population who have a preference for either product and of those who cannot discriminate or have no preference. For the model assumed, the maximum likelihood estimators of the above proportions are derived, their variance-covariance matrix is obtained, and a way of testing the appropriateness of the model is indicated. A decision theoretical formulation is suggested. (Received March 8, 1957; revised March 12, 1957.)

2. Factorial Treatments in Group Divisible Incomplete Block Designs, CLYDE Y. KRAMER AND RALPH A. BRADLEY, Virginia Polytechnic Institute.

Methods of incorporating factorial treatment combinations in group divisible incomplete block designs are given. The factorial treatment combinations are so matched with the basic treatments in the association matrices of the designs that the sums of squares for the factorial effects can be obtained as functions of the original treatment estimators. It is shown first how a two-factor factorial may be incorporated into group divisible incomplete block designs. Single degree of freedom contrasts are obtained for the effects in much the usual way as for factorials in complete block designs. Multifactor factorials and partial factorials are discussed, and a method of obtaining estimates and tests of significance of the effects is given. (Received March 18, 1957.)

3. Iterative Experimentation, G. E. P. BOX, Statistical Techniques Group, Princeton University.

Scientific research is usually an iterative process. The cycle: conjecture–design–experiment–analysis leads to a new cycle of conjecture–design–experiment–analysis, and so on. It is helpful to keep this picture of the experimental method in mind when considering statistical problems. Although this cycle is repeated many times during an investigation,

the experimental environment in which it is employed and the techniques appropriate for design and analysis tend to change as the investigation proceeds. Broadly speaking, one or more of the following four phases can be detected in most investigations: (*a*) a screening phase in which an attempt is made to isolate the important variables; (*b*) a descriptive phase in which the effects of the variables and the positions of interesting regions of the space of the variables are empirically determined; (*c*) a phase leading from (*b*) to (*d*); (*d*) a theoretical phase in which an attempt is made to understand the actual mechanism of the process studied. The roles which statistical methods should properly play in assisting the iterative process at these various phases of experimentation were briefly discussed. (Received March 14, 1957.)

(Abstracts of papers for the Atlantic City meeting of the Institute, September 10–13, 1957)

4. On the Joint Estimation of the Spectra, Cospectrum and Quadrature Spectrum of a Two-dimensional Stationary Gaussian Process, NATHANIEL ROY GOODMAN, New York University (introduced by Leon H. Herbach).

The probability structure of a real two-dimensional stationary (zero mean) Gaussian process $[x(t), y(t)]$, $-\infty < t < \infty$, is specified (in the absolutely continuous case) by $f_{xx}(\lambda)$, $f_{yy}(\lambda)$ the spectral densities of the $x(t)$ and $y(t)$ processes respectively, $c(\lambda)$ the cospectral density, and $q(\lambda)$ the quadrature spectral density $(-\infty < \lambda < \infty)$. The paper treats the problem of jointly estimating (in a suitable sense) $f_{xx}(\lambda), f_{yy}(\lambda), c(\lambda), q(\lambda)$ from a finite part of a sample function of the $[x(t), y(t)]$, $-\infty < t < \infty$, process. An approximation to the joint sampling distribution of the estimators for $f_{xx}(\lambda), f_{yy}(\lambda), c(\lambda), q(\lambda)$ is obtained. This approximate sampling distribution, termed a complex Wishart distribution, serves as the starting point in the derivation of approximate sampling distributions of estimators for functions of $f_{xx}(\lambda), f_{yy}(\lambda), c(\lambda), q(\lambda)$. The paper was motivated by the need of experimenters in fields such as micrometeorology, oceanography, electrical engineering, and aeronautical engineering to statistically estimate "parameters" characterizing their particular physical systems and to treat the sampling variability of estimators for the "parameters." In many cases the "parameters" to be estimated are functions of the densities $f_{xx}(\lambda), f_{yy}(\lambda), c(\lambda), q(\lambda)$ of a real two-dimensional stationary (zero mean) Gaussian process. (Received March 29, 1957.)

5. The Significance Probability of the Smirnov Two-Sample Test, J. L. HODGES, JR., University of California, (By Title).

The approximation $\bar{P}$ for the significance probability P of Smirnov's two-sample test, based on the asymptotic theorem published by Smirnov in 1939, has been shown (Drion) to be accurate when the sample sizes m and n are equal. For this case the error of $\bar{P}$ is of order $1/n$ (Gnedenko). Example: $m = n = 12$, $P = 0.032$, $\bar{P} = 0.034$. A small numerical investigation shows that the accuracy is much poorer when $m \neq n$. Example: $n = 12$, $m = 13$, $P = 0.024$, $\bar{P} = 0.054$. An asymptotic theory is developed for $n \to \infty$ with $m - n > 0$ bounded. The error of $\bar{P}$ is now of order $1/\sqrt{n}$, and to this order is oscillatory as a function of P. This implies that no expression of the simple kind recently published by Korolyuk can be correct. An auxiliary table makes easy the computation of accurate estimates for P when $m - n$ is small and $m \leqq 30$. (Received April 8, 1957).

6. Nonparametric Mean and Variance Estimation from Truncated Data, JOHN E. WALSH, Lockheed Aircraft Corporation.

This paper considers situations where a known number of the smallest values of a sample and a known number of the largest values have been truncated. The problem is to

obtain an estimate of the population mean, an estimate of the standard deviation of this estimate of the mean, and an estimate of the population standard deviation. This paper derives a nonparametric estimate for each of these three cases. These estimates are approximately valid for most continuous statistical populations of practical interest when a small number of sample values are truncated and the sample size is not too small. The mean estimate consists of a linear function of the ordered values of the truncated sample, while each standard deviation estimate is the square root of a quadratic function of these observations. (Received April 10, 1957.)

7. Distinguishability of Sets of Distributions (The Case of Independent and Identically Distributed Chance Variables.), W. HOEFFDING, University of North Carolina, and J. WOLFOWITZ, Cornell University.

Let $\mathfrak{I}$ be a class of tests, based on a sequence of independent chance variables with the common distribution F (assumed to belong to a set $\mathfrak{F}$ of distributions), for testing whether F belongs to one of two disjoint subsets, $\mathfrak{G}$ and $\mathfrak{C}$, of $\mathfrak{F}$. We consider the cases where $\mathfrak{I}$ is either the class of all tests which terminate with probability one if $F \in \mathfrak{F}$, or the class of all fixed sample size tests, or one of several classes intermediate between these two. The sets $\mathfrak{G}$ and $\mathfrak{C}$ are said to be distinguishable in $\mathfrak{I}$ if, for every $\epsilon > 0$, there exists a test in $\mathfrak{I}$ such that the error probability is $< \epsilon$ for all $F \in \mathfrak{G} \cup \mathfrak{C}$. It is shown that if there exists a test in $\mathfrak{I}$ such that the sum of the maximum error probability in $\mathfrak{G}$ and the maximum error probability in $\mathfrak{C}$ is less than 1, then $\mathfrak{G}$ and $\mathfrak{C}$ are distinguishable in $\mathfrak{I}$. Sufficient conditions and necessary conditions for the distinguishability of two sets are expressed in terms of certain distance functions. Certain simple necessary conditions for distinguishability are found to be also sufficient if the class of distributions is suitably restricted. (Received May 20, 1957.)

8. An Extension of Box's Results on the Use of the F Distribution in Multivariate Analysis, SEYMOUR GEISSER AND SAMUEL W. GREENHOUSE, National Institute of Mental Health.

The mixed model in a 2-way analysis of variance is characterized by fixed classification, e.g. treatments, and a random classification, e.g. plots. Under the usual analysis of variance assumptions the proper error for the fixed effect is the fixed $\times$ random interaction component, and the resulting ratio has the F-distribution. If we have individuals instead of plots as the random component and the treatments are correlated, then Box has shown that one may still use the same F-ratio as before as a test of treatment effects; however, the F-ratio does not have the requisite F-distribution, but it can be shown that it is distributed approximately like an F-distribution but with modified degrees of freedom. Box did this for one group of individuals; the authors have extended the Box technique to g groups of individuals and give the modified F-distribution for the tests of treatment effects and treatment $\times$ group interaction effects. (Received May 24, 1957.)

NEWS AND NOTICES

Readers are invited to submit to the Secretary of the Institute news items of interest

Personal Items

H. R. Bright, formerly Deputy Director of the Human Resources Research Office, George Washington University, is now employed as Specialist, Operations

Research and Synthesis, of the Circuit Protective Devices Department, General Electric Company, Plainville, Conn.

The appointments of Dr. Samuel H. Brooks and Julian T. Anderson to the Scientific Staff of Technical Operations, Incorporated, were announced today by Floyd I. Hill, Associate Director of the research and development firm's West Coast office.

Loudon Campbell is now with the Advertising Department of Eaton Laboratories, Norwich, New York.

Z. T. Chang, who has recently come to the United States, has joined Osborne and Thurlow, 39 Broadway, New York City.

Alan Constantine has joined a theoretical research group at the University of Adelaide, Adelaide, Sth. Australia.

Norman R. Garner has resigned as Consulting Statistician to the Quality Control Department of the Naval Powder Factory, Indian Head, Maryland and has joined the Reliability Control Staff of the Aerojet-General Corporation, Azusa, California.

Mark L. Hinkle, Jr. is presently with Western Electric Company as a Development Engineer at the Hawthorne Works in the Department for Mechanization of Equipment Engineering.

B. S. Kawar has been employed by Remington Rand Univac in St. Paul, Minn. to work with the group on "Programming Research and Development".

Truman L. Kochler, Jr. has left the employ of Sylvania Electric Products to accept a position as an experimental statistician with the Operational Statistics Group of the American Cyanamid Corporation.

Dr. R. G. Laha of the Indian Statistical Institute (Calcutta) has been appointed to the position of a Research Associate in the Mathematics Department of The Catholic University of America. He will do research in probability theory and mathematical statistics.

Dr. Eugene Lukas of the Mathematics Department of the Catholic University of America has been granted a leave in order to permit him to accept an invitation to give several lectures at the Sorbonne (Paris).

Jacob Marschak is Professor of Economics and Research Associate of the Cowles Foundation for Research in Economics, Yale University. He is conducting a project, under a contract with the Office of Naval Research, on Decision-Making under Uncertainty.

Michael A. Martino, Jr. has accepted a position as mathematician at the Knolls Atomic Power Laboratory of the General Electric Company in Schenectady, New York.

Albert Mindlin has been appointed Technical Assistant to the Chief of the Statistics Branch, Bureau of Old Age and Survivors Insurance, U. S. Department of Health, Education, and Welfare.

Sidney I. Neuwirth has assumed the position of Operations Research Consultant at Johnson and Johnson, New Brunswick.

Romuald Slimak has recently been appointed Adjunct Assistant Professor of

Mathematics at the School of Engineering, New York University to teach computer methodology.

Lt. (j.g.) F. Beckley Smith, Jr. is now serving on active duty in the U. S. Navy, stationed in Washington, D. C.

Dr. M. D. Springer of the Naval Ordnance Plant, Indianapolis, has accepted a position as Senior Operations Analyst with Technical Operations, Inc., Fort Monroe, Virginia.

David S. Stoller is now a member of the Technical Staff, Computer Systems Division of the Ramo-Wooldridge Corporation, Los Angeles.

Dr. David E. Van Tijn has accepted a position on the staff of Arthur D. Little, Inc.

Irving Weiss has left the Mathematics Department of Lehigh University to join the Bell Telephone Laboratories, North Andover, Massachusetts.

New Members

The following persons have been elected to membership in the Institute

February 6, 1957 to May 2, 1957

Anderson, Norman H., Ph.D. (Univ. of Wisconsin), Postdoctoral Fellow, Social Science Research Council, *Yale Univ. Department of Psychology, 333 Cedar St., New Haven, Conn.*

Askovitz, S. I., M.D. (Univ. of Penna.), Medical Statistician, Albert Einstein Medical Center, York and Tabor Roads, Phila. 41, Pa.; Chief of Tumor Registry, School of Medicine, Univ. of Penna., Philadelphia 4, Pa.; *4900 North 9th Street, Phila. 41, Pa.*

Belson, Irving, M.A. (Columbia Univ.), Statistician, *Mine Safety Appliances Research Corp., Callery, Pennsylvania.*

Bentley, D. L., B.S. (Stanford Univ.), Student, *Dept. of Statistics, Stanford Univ., Box 2695, Stanford, California.*

Borden, Julien Louis, B.A. (U.C.L.A.), Grad. Student, Research Assistant, U.C.L.A., Los Angeles, California; *6626 W. 5th St., Los Angeles 48, California.*

Brock, Dan A., M.A. (Southern Methodist University), Analyst and Statistician, Dallas City Water Works, City of Dallas, Texas; *2015 Commerce Street, Dallas, Texas.*

Cohen, Arthur, B.A. (Brooklyn College), Student, Dept. of Math., Columbia Univ., New York City, New York; *1925-66th St., Brooklyn 4, New York.*

Doornbos, R., M.S. (Univ. of Groningen), Statistician, Uhilever N.V., Unsemnpark 1, Rotterdam, Netherlands; *Lever's Zeep Mij, N.V., Parallelweg, Vlaardingen, Netherlands.*

Dorff, M. R., M.S. (Carnegie Inst. of Tech.), Student, Statistical Laboratory, Iowa State College, Ames, Iowa; *128 Lynn Ave., Ames, Iowa.*

Fedick, John, M.S. (Univ. of Michigan), Assistant to Treasurer, *Vickers, Inc. Administrative and Engineering Center, P.O. Box 302, Detroit 32, Michigan.*

Galligan, Agnes M., A.B. (Brown Univ.), Analytical Statistician, Quartermaster Research and Development Command, EPRD, Natuk, Mass.; *104 Lanoseer St., West Roxbury, Mass.*

Gnanadesikan, Ramanathan, Ph.D. (Univ. of N. C.), Research Assistant, Univ. of North Carolina, Dept. of Statistics, Chapel Hill, N. C.; *714 E. Franklin St., Chapel Hill, N. C.*

Hans, Otto, Candidate of Physical and Mathematical Sciences (Charles University), Scientific Worker, Czechoslovak Academy of Sciences, Institute of Radio-Engineering and

Electronics, Charles Street 2, Prague 1, Czechoslovakia; *until June 31, 1957, Krusinova 44, Praha 14, Czechoslovakia; from July 1, 1957, Dvorecka 3, Praha 15, Czechoslovakia.*

Horner, Theodore W., Ph.D. (N. C. State College), Asst. Prof., *Statistical Lab., Iowa State College, Ames, Iowa.*

Jensen, Arne, Ph.D. (Univ. of Copenhagen), Actuary, Copenhagen Telephone Co.; *Kraedevej 74, Virum, Denmark.*

Johnson, Eugene A., Ph.D. (Univ. of Minn.), Asst. Prof. of Biostatistics, *School of Public Health, Univ. of Minnesota, Minneapolis 14, Minn.*

Kao, Richard C., Ph.D. (Univ. of Illinois), Asso. Mathematician, *the RAND Corp., Santa Monica, Calif.*

Krane, Scott A., B.A. (Simpson College), Grad. Student, Stat. Laboratory, *Iowa State College, Dept. of Statistics, Ames, Iowa.*

Lilliefors, Hubert W., M.A. (Michigan State Univ.), Mathematician, Operations Evaluation Group, MIT, P.O. Box 2176, Potomac Station, Alexandria, Virginia; *4411 N. Pershing Drive, Arlington, Virginia.*

Makabe, Hajime, B.S., M.A. (Tokyo Inst. of Tech.), *Instructor of Math., Tokyo College of Science, Kagurazaka, Shinjuku-ku, Tokyo, Japan.*

Malik, Azizul, M.A. (Math., Muslin Univ., Aligazh), M.A. (Economics, Agra Univ.), Research Officer and Honorary Lecturer in Mathematical Statistics, Karachi Univ., *Central Statistical Office*, Ministry of Economic Affairs, *Muhammadi House (7th Floor), McLeod Road, Karachi.*

Martino, M. A., Jr., Ph.D. (Univ. of Illinois), Mathematical Analyst, National Security Agency, Washington 25, D. C.; *1412 Patrick Henry Drive, Falls Church, Va.*

Munoz-Febar, Luis E. Marquez, Licenciado en Ciencias Estadisticas (Universidad Central), Assor Tecnico Seccion Estadisticas, Direccion General de Estadistica, Torre Sur. Centro Simon Bolivar, Ministerio de Fomento-Direccion General de Estadistica, Seccion Est. Econ. 5° Piso; *Domicilio Particular, Arismendi a Pichincha 114, San Agustin del Norte, Caracas, Venezuela.*

McCullough, R. S., B.A. (Toronto), Teaching Fellow, Dept. of Math., Univ. of Toronto, Toronto, Ontario; *131 Heath St., East, Toronto, Ontario, Canada.*

Ozanne, Paul, B.S. (Univ. of Wisconsin), Asst. to Director of Planning, *Anheuser Busch, 721 Pestalozzi St., St. Louis 18, Mo.*

Panizzon, G., Ph.D., *Istituto di Statistica, Universita di Padova, Padova, Italy.*

Park, Han Shick, B.S. (College of Educ., Seoul National Univ.), Lecturer, *College of Education, Seoul National University, Dept. of Mathematics, Yongtoo-Dong, Seoul, Korea.*

Parmenter, Ellis F., Ph.D. (Brown Univ.), Sr. Research Engineer, *Statistical Consultant, Champion Paper and Fibre Co., Hamilton, Ohio.*

Sammons, William H., M.S. (Univ. of Kentucky), Dept. of Agriculture, Soil Conservation Service, Central Technical Unit, SCS, Hillculture Bldg., Agricultural Research Center, Beltsville, Maryland; *8714-63rd Ave., Berwyn Heights, Maryland.*

Smith, James D., M.S. (Univ. of Minnesota), Applied Science Representative, IBM, 590 Madison Ave., New York, N. Y.; *1200 Ind Ave., So., Minneapolis, Minn.*

Uzawa, Hirofumi, B.S. (Univ. of Tokyo), Research Associate, *Applied Mathematics and Statistics Laboratory, Stanford University, Stanford, California.*

Wall, Francis J., M.S. (Univ. of Colorado), Mathematician, Remington Rand Div. of Sperry-Rand Corp., *610 East 77th Street, Richfield 23, Minnesota.*

Walsh, Richard R., B.S. (Stanford Univ.), Student, Stanford University, Stanford, California, *1990 Avy Ave., Menlo Park, Calif.*

Wicks, Byron E., A.B. (Univ. of California), Operations Research Assistant, Bank of America, San Francisco, Calif.; *2381 Maricopa Ave., Richmond, Calif.*

Woodson, G. Stanley, M.A. (Univ. of Denver), Chief, Biostatistics Section, Medical Research Directorate, Chemical Corps, U. S. Army Chemical Center, Maryland; *111 South Reed Street, Bel Air, Maryland.*

AUSTRALIAN MATHEMATICAL SOCIETY

During 1956 at a meeting in Melbourne 121 pure and applied mathematicians formed the Australian Mathematical Society. About a fourth of these are professional statisticians and the 1956 meeting featured a number of statistics sessions. The Council consists of T. M. Cherry (President), C. S. Davis (Treasurer), T. G. Room (Public. Secr.), A. L. Blakers, H. S. Green, H. O. Lancaster, H. C. Levey, P. A. Moran, E. J. G. Pitman, A. H. Pollard, L. C. Woods. The General Secretary is J. P. Ryan (Math. Dept., Univ. of Melbourne, Carlton N. 3, Victoria, Australia). The next meeting will be held in Sydney on 28–31 August 1957.

PUBLICATIONS RECEIVED

COWDEN, D. J., *Statistical Methods in Quality Control*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1957, xxiv + 727 pp., $9.00.

DIXON, W. J., and F. J. MASSEY, JR., *Introduction to Statistical Analysis*, McGraw-Hill Book Company, Inc., New York, 1957, xiii + 488 pp., $6.00.

OWEN, D. B., *The Bivariate Normal Probability Distribution*, Office of Technical Services, Department of Commerce, Washington, D. C., 1957, 136 pp., $.65.

RESNIKOFF, G. J. and G. J. LIEBERMAN, *Tables of the Non-Central t-Distribution*, Stanford University Press, Stanford, California, 1957, 389 pp., $12.50.

BIOMETRIKA

Volume 44 **Contents** **Parts 1 and 2, June 1957**

The subscription price, *payable in advance*, is 45s. inland, 54s. export (per volume including postage). Cheques should be drawn to Biometrika and sent to "The Secretary, Biometrika Office, Department of Statistics, University College, London, W.C. 1." All foreign cheques must be in sterling and drawn on a bank having a London agency.

ECONOMETRICA

Journal of the Econometric Society

Contents of Vol. 25, No. 4 - October, 1957

JOURNAL OF THE

ROYAL STATISTICAL SOCIETY

Series B (Methodological)

Vol. XIX, No. 1, 1957

The Royal Statistical Society, 21, Bentinck Street, London, W. 1

SKANDINAVISK AKTUARIETIDSKRIFT

1956 – Parts 1 – 2

Contents

Annual subscription: $5.00 per year

Inquiries and orders may be addressed to the Editor,

GRANHÄLLSVÄGEN 35, STOCKSUND, SWEDEN